INVITED PAPER

Event Mining in Multimedia Streams

Research on identifying and analyzing events and activities in media collections had led to new technologies and systems.

By Lexing Xie, Hari Sundaram, and Murray Campbell

ABSTRACT | Events are real-world occurrences that unfold over space and time. Event mining from multimedia streams improves the access and reuse of large media collections, and it has been an active area of research with notable recent progress. This paper contains a survey on the problems and solutions in event mining, approached from three aspects: event description, event-modeling components, and current event mining systems. We present a general characterization of multimedia events, motivated by the maxim of five "W"s and one "H" for reporting real-world events in journalism: when, where, who, what, why, and how. We discuss the causes for semantic variability in real-world descriptions, including multilevel event semantics, implicit semantics facets, and the influence of context. We discuss five main aspects of an event detection system. These aspects are: the variants of tasks and event definitions that constrain system design, the media capture setup that collectively define the available data and necessary domain assumptions, the feature extraction step that converts the captured data into perceptually significant numeric or symbolic forms, statistical models that map the feature representations to richer semantic descriptions, and applications that use event metadata to help in different information-seeking tasks. We review current event-mining systems in detail, grouping them by the problem formulations and approaches. The review includes detection of events and actions in one or more continuous sequences, events in edited video streams, unsupervised event discovery, events in a collection of media objects, and a discussion on ongoing benchmark activities. These problems span a wide range of multimedia domains such as surveillance, meetings, broadcast news, sports, documentary, and films, as well as personal and online media collections. We conclude this survey with a brief outlook on open research directions.

KEYWORDS | Data mining; events; indexing; multimedia; pattern recognition; review; survey

Manuscript received July 11, 2007; revised December 3, 2007.
L. Xie and **M. Campbell** are with the IBM T.J. Watson Research Center, Hawthorne, NY 10532 USA (e-mail: xlx@us.ibm.com; mcam@us.ibm.com).
H. Sundaram is with Arts and Media Engineering, Arizona State University, Tempe, AZ 85281 USA (e-mail: hari.sundaram@asu.edu).
Digital Object Identifier: 10.1109/JPROC.2008.916362

I. INTRODUCTION

Events can be defined as real-world occurrences that unfold over space and time. In other words, an event has a duration, occurs in a specific place, and typically will involve certain change of state. Using this definition, "a walk on the beach," "the hurricane of 2005," and "a trip to Santa Barbara" would all qualify as events. Events are useful because they help us make sense of the world around us by helping to recollect real-world experiences (e.g., university commencement 2006), by explaining phenomena that we observe (e.g., the annual journey of migrating birds), or by assisting us in predicting future events (e.g., the outcome of a tennis match).

While no statistics are available on how much real-world event content is being captured in multimedia, we can infer its scale from the fact that video already accounts for more than half of internet traffic, with YouTube alone taking 10% [114]. The increasing degree to which real-world events are captured in multimedia further enhances our ability to not only archive events, but to also recollect, reason about, and relate to other events. For instance, analyzing the video of traffic patterns on a highway can help plan construction and reduce congestion, and pinpointing anomalies from closed-circuit surveillance video may prevent small crimes or terrorist attacks. With current technologies, however, there is little or no metadata associated with events captured in multimedia, making it very difficult to search through a large collection to find instances of a particular pattern or event. There is a clear need for automated analysis of multimedia streams to improve the use and re-use of multimedia repository.

This paper surveys event mining, and we cover aspects in describing, modeling, and analyzing events from multimedia. Event description is the basis for setting up event detection problems and defines goals of such systems. Our

survey on event modeling decomposes the media analysis problem and discusses the main issues and common solutions for each component.

Event description tries to structure the semantics of real-world event representations in order to clarify assumptions and assist in event analysis. We draw upon the principle of the five "W"s and one "H" (5W1H) in journalism. We present examples of these six dimensions (*who*, *when*, *where*, *what*, *why*, and *how*) along which events can be described. We also discuss reasons for semantic variability, i.e., how there are multiple possible descriptions of an event, and why this should happen. The semantic variability has two aspects: event aggregates and implicit event facets. There are two main reasons for this variability: context and the sensory gap between real-world happenings and what are captured in media.

Event detection is the process of mapping multimedia streams to event descriptions. We examine the five major components of an event-modeling system: target properties, data capture, feature representation, computational model, and application. The target event properties, data capture, and application components define the problem, while the feature representation and computational model make up the solution.

A significant portion of this survey is devoted to reviewing a range of existing and ongoing work on event detection, put in the perspectives of the faceted event descriptions and the event-modeling components. We group event detection systems based on their problem setup and detection target. These include: detecting known events and activities from one continuous capture, such as in surveillance applications; unsupervised event detection and pattern detection, such as detecting routine but unknown behaviors in videos of meetings; or recovering event semantics from collections of photos and videos. We also review current benchmarks that are related to event detection, which provides a perspective for both practitioners and researchers about several consensus problems in the community along with relative performance comparisons.

In this paper, we intend to clarify and categorize three facets of the event mining problem: target semantics, the problems, and the solutions. The main contributions of this paper are as follows.

1) In event description, we rely on the 5W1H framework and build on prior work on event representation [144] and interactive annotation [142]. We reorganized the earlier frameworks so that the semantics of the aspects are aligned with common sense and more complete (see Section II-A7). We also discuss how the embodiment in real-world media influences event aspects and event detection. In particular, we explain the semantic variations that appear in each aspect and discuss how intuitive aggregates and media production can affect different event facets, e.g., the notion of media time and real-world time in *when*, and semantic aggregations in *what*, *where*, and *when*.
2) We provide a dual overview of the components and systems for event detection. In terms of coverage, there have been a few surveys in multimedia analysis and indexing but none have discussed events in depth. There are surveys on extracting generic static semantics [151], in which event is an important but distinct subset mentioned in the passing, or on image/video retrieval [20], [118], [129], [131], [155], which can be an application for event analysis. In terms of organization, Section II provides a component overview similar to a few image/video retrieval surveys [78], [129], [131], [155]; Section IV surveys existing event detection systems grouped by common problem scope and solution components, similar to earlier system surveys [20], [118]. This dual structure is necessary since event-modeling problems are very diverse, and there are a number of problem setups and components that the community is actively working on.

Modeling event semantics and extracting them from multimedia data are of interest to numerous areas outside the multimedia analysis community. This is due to a number of reasons: First, the understanding of event semantics in real-world media collections draws upon a number of research areas that are traditionally separate: knowledge representation [93], [122], computer vision, auditory analysis, natural language processing, machine learning, databases, to name a few. Problems in this area provide synergies among these areas for the understanding of multimedia content and present new challenges such as multimodal fusion and learning with structured input–output. Second, with the ubiquitous presence of multimedia data in our lives for both informational and entertainment needs, better understanding and modeling events will enable better user experiences and improve system design in closely related areas such as computer–human interaction, multimedia communication and transmission, multimedia authoring, etc. Finally, the underlying data processing and learning methodologies used here are very similar to those seen in many other domains—stream data mining, network measurement and diagnosis, bio-informatics, business processing mining, and so on.

We limit the scope of this paper in three main ways. First, our main focus is on event analysis from existing archives and repositories. While event analysis can certainly benefit from systematic representation and capture of metadata [144], solving the problem with incomplete data is more likely to be valuable for many real-world applications and existing content collections. Second, although we provide a framework for thinking about the problems and solutions in the event and pattern mining space, our review of specific approaches is by no means complete, in part due to the rapid development in the area. And finally, while our focus is on

the general problem of event modeling across multiple modalities, the majority of the work reviewed here relies to a significant extent on visual information, and less on other modalities such as sound, freeform text, or structured metatags. This is in part due to the perspective of the authors and also due to the relatively limited work already done on the analysis across the different modalities.

The rest of the paper is organized as follows: Section II presents event characterization along six commonsense dimensions and discusses their semantic variability. Section III examines the event-modeling problems and presents five major components of a event-modeling system. Section IV reviews the state of the art on event detection and modeling. Section V concludes the paper with a brief outlook.

II. EVENT DESCRIPTION

Events can be described in many legitimate forms, such as a walk on the beach, the hurricane of 2005, a trip to Santa Barbara. Real-world events are captured with images, sounds, videos, and text, these media and the underlying semantic facets such as who, where, when, and what support the understanding of events. Diverse as they seem, there are underlying structures in event semantic that allow a systematic organization of these descriptions. In this section, we resort to six common-sense aspects called 5W1H to characterize events in multimedia. We will also discuss factors that account for the variability in these semantics.

A. 5W1H of Multimedia Events

We refer to journalistic practices for covering real-world events to design a systematic way of describing events. A key maxim in journalism is to use the six interrogatives—*who?*, *when?*, *where?*, *what?*, *why?*, and *how?* to develop a comprehensive reportage of the event. We adopt the same six facets to describe events in multimedia streams, because they are key semantic attributes that are sufficient and necessary to summarize an event, as prescribed by journalism principles, and also because they can be preserved in the process of capturing a real-world event into multimedia sequences, as can be seen in the example that follows.

Fig. 1 shows an example of a real-world event: 2007 NBA finals, Game 4. This event can be summarized along the six facets—*who*: the Cleveland Cavaliers and the San Antonio Spurs; *when*: June 14th, 2007; *where*: the Quicken Loans Arena, Cleveland, Ohio, U.S.; *what*: the Cavaliers play the Spurs in Game 4; *how*: the Spurs win the NBA Finals 4-0 with a 83-82 victory in this game; *why*: the Spurs exhibited teamwork and played good defense. A human observer, not surprisingly, can recover the same summary from either snapshot of the game: a news article or a TV broadcast of the game.

These event facets, referred to as 5W1H for short, are types of metadata, or structured, descriptive information about a multimedia clip. Similar to the distinction made in library resources, we distinguish the metadata in multimedia into two broad categories: *intrinsic* and *extrinsic* metadata [127]. Being *intrinsic* means being attached to the multimedia clip during capture and production, and remaining relevant to the resource it describes, no matter what, the context. Examples of *intrinsic metadata* include the widely used EXchange Image File (EXIF) [69] metadata, the media format, bitrate, or the viewers/players/platform on which the multimedia clip can be viewed or edited. *Extrinsic metadata* include additional information that describes the media clip, that helps present its meanings within a context. The 5W1H are part of the *extrinsic* metadata of the media content, since they serve to describe semantics rather than signal-level composition of the media; their interpretations may be variable across different context, as will be shown by the examples in Sections II-A1–A6 and the discussions in Section II-B3. Moreover, they are actually *invariant* to certain *intrinsic* information of the multimedia clip given the same context, e.g., the same scene, taken by two different cameras from exactly the same viewpoint and camera settings can share the same set of semantic descriptions. We are interested in the 5W1H, as they are extrinsic value-adds to media representation, that can both enrich the representation and help event-based media applications and information seeking tasks, as discussed in Section III-B.

We now go on to discuss each of the six event facets in more detail: how they help describe an event, how their scope, meanings, and applicability change when applied to multimedia data, and how new values can be defined and created at different stages of media production and consumption. We also compare and connect the 5W1H with the work on common event representation model [144], from which this work draws upon.

1) When—Time: Time is one of the key components of an event, yet the description of time can take many forms. Consider the event of "my conference trip to ACM Multimedia"; it is possible to specify the time attribute in a variety of ways—"October 21st–25th, 2006" (exact), "before Thanksgiving" (relative), "every year" (periodic), "in the fall" (at a coarse granularity), "in the best season of the year" (affective).

Importantly, there is more than one temporal coordinate in events captured by media: real-world time, media time, time of post-processing, time of uploading and sharing on a website, time that comments are posted, etc. The relationship between real-world time and media time is the most important one, since this relationship can change the semantics of the media components (e.g., shot sequences) as explained in the following. Furthermore, this relationship is set after the media is produced and is invariant to how media is distributed or consumed.

The real-world time refers to the absolute, unambiguous time that an event takes place in the physical world, e.g., 7:00 pm Pacific Time, October 22nd, 2006. Media

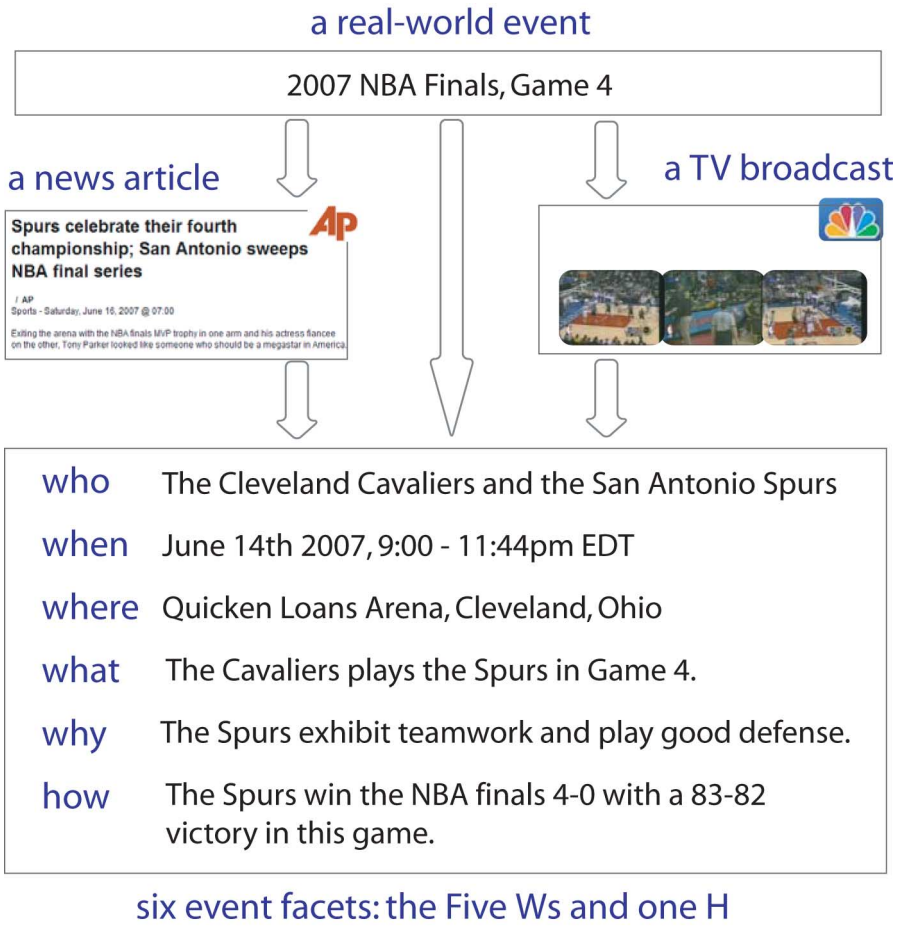

Fig. 1. ***Example of a real-world event—Game 4 of 2007 NBA Finals—summarized into six semantic facets from journalism coverage (middle left) and multimedia clip (middle right).***

time refers to the relative value within the media stream, e.g., 10 minutes after the news began, third shot in the film, the first picture in a vacation collection, etc. Note that media time is observable from the data streams, while real-world time can either be known, e.g., captured in media metadata such as EXIF, or it can be hidden or imprecise, e.g., the interview shot at 10 minutes into the news shot may have been taken sometime during the day that is unknown to the viewers. Reconciliation of real-world clips that refer to the same event, or alignment of media time and real-world time, can become a challenging problem in the absence of additional temporal metadata.

Fig. 2 illustrates possible mappings between real-world time and media-time coordinates with four example content domains. Fig. 2(a) represents a continuous media capture, such as surveillance videos and meeting captures.

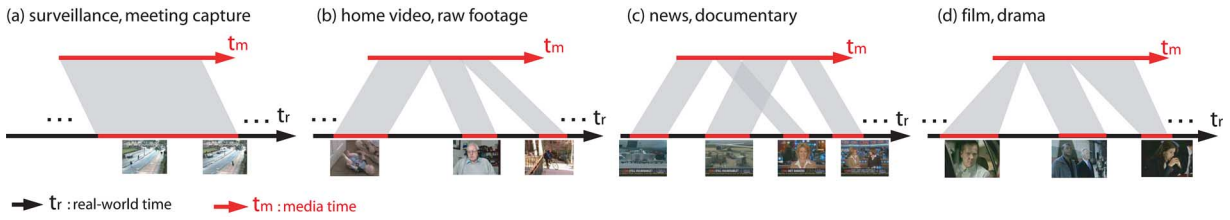

Fig. 2. ***Four possible mappings between real-world time and media time for different types of real-world media clips: (a) surveillance, meeting capture; (b) home video, raw footage; (c) news, documentary; and (d) film, drama.***

Media in such domains is typically one long shot,[1] where there is one-to-one correspondence with the media time and real-world time through a constant offset. This setup easily generalizes to a multicamera continuous capture, where several capturing devices are time synchronized. Fig. 2(b) represents intermittent sequential media capture, such as home videos or raw professional footage. In such capture, the media clip typically consists of multiple shots, each of which has a different temporal shift with respect to the real-world time, and the temporal order of the shots preserve the temporal precedence in the real world. Fig. 2(c) represents edited footage with intermittent capture and temporal reordering. This can happen in broadcast news and documentaries, where different shots not only have different shift with respect to the real-world time, the temporal precedence between shots are also reordered among the field footage, interviews, and studio shots. Finally, Fig. 2(d) represents edited video with reordering and temporal scaling. This can happen in film, sports replays, and other drama genres. Such videos may not map to specific real-world times, they may also employ special camera and cinesthetic techniques [124] to *compress* or *expand* time in a shot.

The relationship between real-world time and media time varies widely across different content domains. This poses a great challenge for event recognition systems to generalize across domains, since the assumptions about such relationships are implicitly coded in the system and are rarely explicit and reconfigurable in the metadata. Many event recognition systems do not yet address the issue of recovery between media time and real-world event time. This has not yet been a critical problem as current recognition systems have been focused on recognizing the objects/people in the given content domain. However, this issue will become increasingly important when researchers begin to address the increasing need of retrieving media stream from diverse sources related to events (e.g., finding "a kid's birthday party" on YouTube) [75], [76].

2) Where—Location: Space is another key multimedia variable that can be used to index and interpret events. Prior work [94] reveals that time and location are the most important attributes for people to recall real-world events. Similar to time, the description of location can also take many forms, e.g., "500 W. 120th St., New York, NY 10027" (absolute and exact); "five miles northeast" (relative); "not far from the Hudson River" (approximate). It can also be used at different granularities, "the seventeenth floor auditorium in the GE Building" or "in New York City."

Space, like time, is used in two coordinate systems—the absolute spatial location where an event occurs and the display space where creators can reorganize elements to communicate a specific affect or meaning. The relationship between absolute real-world event locations and their corresponding media locations is not a straightforward mapping, in a similar manner as that for time (Fig. 2). Changes in the real-world event location are not necessarily reflected in changes to an object or a person's location in the video. Geo-spatial visualization of media [9], [11] is a possible way to visualize event location changes in correspondence to media location changes. We note that in creative domains, there is very little in relationships between real-world event locations and how they are manifest on screen. Film makers routinely alter our perception of space (as well as time) through clever event capture and event editing [17], [32].

3) Who—Subject: The *who* field has typically referred to the subject in the media clip—*who is in the photo/video?* However, this can quickly get complex given the entire media processing chain. For example, one could ask *who took the photo*, *who edited the photo*, *who posted the photo online*, *who has seen the photo*, etc. While the first two questions are directly connected to the event itself—event participation and event capture, the last three questions are about operations on the media clip that remediates the original event—media editing, communication, and viewing. These different attribute values are useful in different contexts. For example, *who* edits the media clip may be important in the context of a media production house, where the "event" refers to an edit of the raw capture, not the original event captured in the content. It may be important to be able to retrieve the media clips for the editing event based on the media clip editor.

Similar to *when* and *where*, the *who* facet in event participation can have many forms, e.g., "Abraham Lincoln the U.S. President" (absolute), "a politician" (generic category), "a group of suited people" (a collection), "Ph.D. students more senior than Tom Sawyer" (relative), etc. Sometimes the *who* facet need not refer to people or impersonated characters, it may simply be the subject of action, e.g., "rocket launching," or change of state, e.g., "earthquake," "hurricane."

4) What—Actions, Activities, and Their Aggregates: The *what* field describes the change or action taking place in the media clip. It answers the question *what is happening in this clip?* For example, in a video clip containing a stroll by the beach, the what field would be described as "walking," "stroll." The answer also depends on the required granularity or levels of abstraction. For example, the *what* question could be alternatively answered as—*Tom walking on Venice beach* (highly specific), to *a person walking* (abstract).

In addition, the *what* question can be answered in different ways, depending on the event that the clip belongs to. For instance, we could ask about the action, *what is Bob doing*; the object, *what is Bob working on*; the

[1]In this paper, a shot is just one single continuous camera take. This is a mechanical definition—equivalent to when the camera is started, to when it is switched off.

goal *for what is Bob preparing the presentation for*. The utility of the questions depend on the event and the specific user context.

5) Why—Event Context: The *why* field provides reasons for an event to happen, e.g., "why did Tom's party take place?" This question cannot be answered by examining a single event in isolation. In this case, the reason why the party took place could depend on another event—"he successfully defended his thesis." The set of events that are needed to understand the semantics of a specific event form the context of the event. Note that this is different from the event description context. The event description context is the set of conditions (Section II-B3) that affect the *values* of the 5W1H attributes not the set of other events that lead to, or result from, this event.

6) How—Event Dynamics: The *how* field answers a subtly different question from *why* or *what*. It is the answer to the questions *how was the event?* or *how did this event come about?* In the preceding example, the *how* aspect to Tom's event can be "it was a fun party," "forty people came and congratulated him," "the guest lists was determined three weeks ago, and a few friends helped Tom with the shopping and cooking since the day before," "the party went way over budget." The *how* facet helps understand the event dynamics, it can either modify the *what* facet from different angles, or like the *why* field connects to other related events and can only be answered in context.

We note that the *why* and the *how* attributes of an event may depend on other events. Most current multimedia analysis research does not try to answer these questions from a single media instance. *Why* and *how* is part of knowledge representation and reasoning in core AI, and their extraction and inference can benefit from work in this area such as event calculus [93]. Also note that each of the six facets can take multiple values, Section II-B discusses the reasons for this semantic variability from three viewpoints.

7) Connections With Earlier Event Representation Models [142], [144]: The event description through the 5W1H draws upon the work on event models by Westermann and Jain [144], or earlier work by Vendrig and Worring on film annotation [142].

Westermann and Jain proposed six aspects for event representation, to which four out of our six event facets are similar, i.e., the *temporal*, *spatial*, *informational*, and *causal* aspects there correspond to *when*, *where*, *who*, and *why* attributes, respectively. It is worth noting that their focus is on event capture and representation, while our primary concern is on adding semantic descriptions by event analysis. Their representation also included the intrinsic media metadata (modalities, media format, size, etc.; see Section II-A) as the *experiential* aspects, and event–subevent relationships in the *structural* aspect. In event analysis, we are often working with a *given* set of media clips for which the intrinsic metadata are already *fixed* (e.g., the NBA game broadcast stream was already captured). Moreover, we find that the event–subevent relationship can be inferred through continuity or similarity relationships in the 5W1H, the hierarchical nature of events (Section II-B1), as well as reasoning in knowledge representation [93].

Vendrig and Worring used the four aspects also common in Westermann and Jain: *who*, *when*, *where*, *what*. In addition, we include the *what* and *how* attributes for the semantic tags and modifiers associated with the event. These attributes often cannot be recovered from the media content and other events aspects alone since they values will vary with respect to context and task (Section II-B), yet users tag using these semantics [10], and they are becoming increasingly relevant in tagging, annotation, and content analysis tasks.

B. Why are Semantics Variable?

We have seen a significant amount of variation in each of the key event facets. This variation mainly falls into two types: 1) varying semantic granularity and 2) implicit or hidden semantics. Furthermore, these variations come from two causes: the dependence on context and a multimedia clip as an incomplete capture of everyday experiences.

1) Multilevel Semantics and Event Polysemy: A media clip can belong to many different events due to the different granularities at which the clip is described.

Fig. 3 illustrates event polysemy in a conference scenario: A video clip can be either described as "Dr. A runs live demo of system X in Rm IA, 11:20 am on Oct 24th" (Box e), or "Dr. A and Ms B present papers at the ACM Multimedia conference content analysis session, 10:20–11:50 am, Oct 24th" (Box d). While both descriptions are valid, the second description applies to a longer time span and can be regarded as containing the first description as a subevent. Furthermore, the same clip can also be generally described as part of "ACM Multimedia conference technical presentations" (Box b), as opposed to "the social and leisure activities at ACM Multmedia" (Box c). Through this example, we can see that event polysemy results from aggregating or expanding event descriptions in one or more facets. Also, (d) can be derived from (e) and a few other events instances (not shown) that share the same location (*where*) and are adjacent in time (*when*); while "conference lunch" (Box f) and "a walk on the beach" (Box g) can be aggregated since their *what* facets can both be described as "social and leisure activity."

Fig. 4 further illustrates semantic aggregation along specific facets. Without loss of generality, we plot time (the *when* facet) as the x axis and location (the where facet) in the y axis, and we use colored squares to represent technical seminars on different topics (the *what* facet). Fig. 4(a) shows

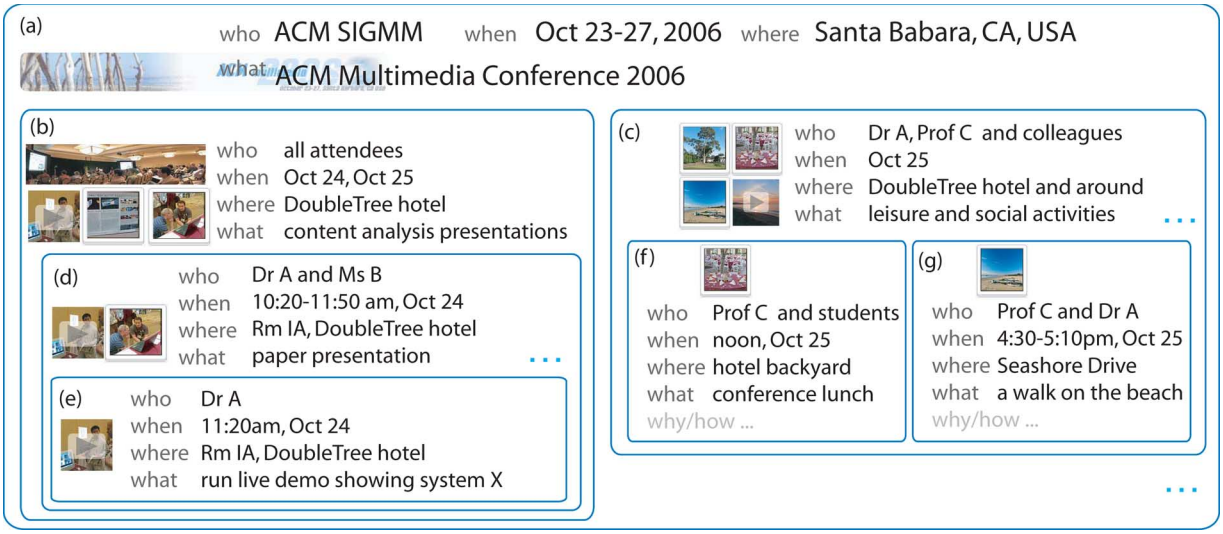

Fig. 3. ***Multilevel event semantics by aggregating event instances: ACM Multimedia Conference (see Section II-B1). Photos courtesy of ACMMM06 Flickr group: http://www.flickr.com/groups/acmmm2006/.***

that talks that are adjacent in time and location can be grouped into one aggregate event, such as "ACM Multimedia technical sessions"; Fig. 4(b) shows that seminars that share the same location and regular temporal intervals belong to one semantic theme "multimedia research group weekly seminars"; Fig. 4(c) shows that a set of seminars that took place in diverse time and locations actually share the author (the *who* facet) and topic (the *what* facet), summarized as "MyLifeBits talk by Gordon Bell."

From the examples above, we can see that event polysemy results from the aggregation of event instances along one or more of the semantic facets among the 5W1H. There are many meaningful aggregating operations, such as: continuity [Fig. 4(a)], shared values [Fig. 4(c)], or regular intervals [Fig. 4(b)]. In addition, event polysemy is the effect of multilevel semantics within one event, while semantics is certainly affected by contextual factors such as the cause, effect, task, as will be discussed in Section II-B3.

2) Implicit Semantics and Hidden Facets: Aside from the plurality of valid meanings, event facets can also be implicit or unknown.

Implicit facets can be recovered due to their inherent correlation with the known facet. For instance, we can describe Box (c) in Fig. 3 as "social and leisure activities during ACM Multimedia 2006," with only the aggregated *what* facet—from this description, we can recover the *when* and *where* facets from "ACM Multimedia 2006," assuming access to a corresponding knowledge base.

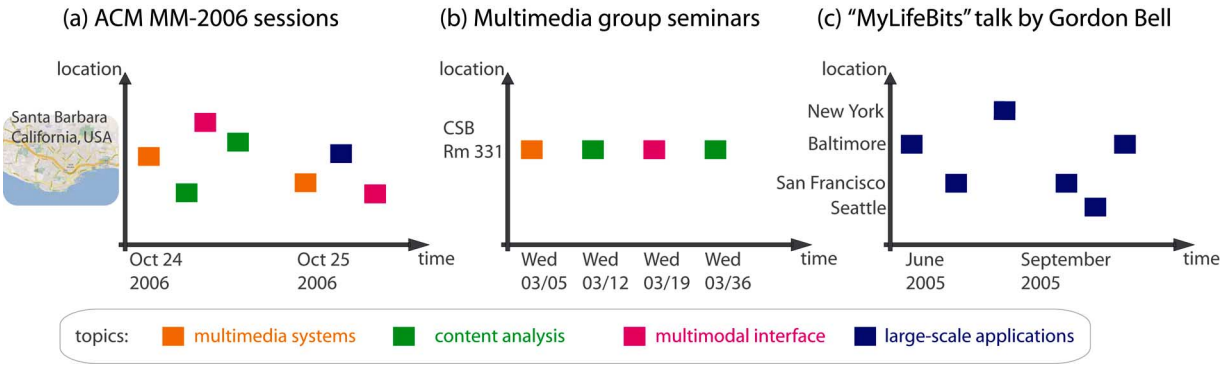

Fig. 4. ***Example event aggregations over different semantic attributes. (a) ACM MM 2006 conference sessions, continuous over space and time; (b) multimedia group seminars, at a constant location and regular intervals over time; (c) topic-specific seminars "MyLifeBits" talk by Gordon Bell, diverse in location or time.***

Unknown facets result from either imprecise correlations or the incompleteness in event capture and annotation. For example, a news story covering an important initial public offering (IPO) includes an interview with an economist scholar. While the theme of the interview is tied with the news story, the exact *when* and *where* facets are unknown—the time should be after the public announcement of this news and before the TV broadcast is aired, but the exact real-world time remain unknown due to the imprecise specifications from the parent event. The interview clip may contain a head-and-shoulder indoor shot of the interviewee, there the location is neither mentioned nor discerned from the video itself.

While we acknowledge the polysemy and scarcity of semantic event descriptions, such imprecision is often seen in real applications without affecting our use of such descriptions. For example, aggregating shared *what* facets among event instances is of interest to many current event detection systems (Section IV). This is because problems in visual event detection are decomposed as recognizing occurrences of similar actions or activities—commonly referred to as *event class* detection. This problem can include, for example, *person walking*, *baseball homerun*, with instances occurring in different times and/or locations.

There are two reasons that caused imprecise event descriptions in practice: 1) Event context that narrows down the values of event facets or supplies necessary values for the implicit facets (Section II-B3). 2) The sensory gap between everyday experience and captured media clips, responsible for the loss of unknown values in media production process (Section II-B4).

3) What is the Role of Context?: Event context refers to the set of interrelated multisensory conditions that affect the choice of event facet values.

Prior work has studied context in ubiquitous computing and media retrieval applications. Dey [36], [37] has built context-sensitive ubiquitous computing systems that take into account location, identity, activity, and time. There, context is defined as [36] "any information that can be used to characterize the situation of an entity, where an entity is a person, place or object that is considered relevant to the interaction between a user and an application, including the user and application themselves." Winograd [145] bounded the definition of context as the set of information that is relevant for the current communication. Note that these definitions imply to two important properties—1) Context is a dynamic construct [38], [51] and 2) Context is related to knowledge and cannot be discussed independent of it [106]. Context has also been used in multimedia content analysis applications. Mani and Sundaram [87] construct a graph-based context representation that helped retrieval of a large collection of personal photos.

For event description, context includes knowledge, user task, and event history. The use of context can help disambiguate event descriptions by supplying hidden event facet values, narrowing down and deciding the angle from which the questions about the 5W1H are asked. Knowledge is an emergent set of multisensory facts, a subset of which is in attention and affects the attributes that describe an event. For example, the background knowledge about soccer matches can help fill in the location for "Italy won FIFA 2006" as "Berlin, Germany." A user may choose to describe a picture as "the golden shot that decided FIFA 2006" if his/her task is to annotate a set of FIFA 2006 photos, while the same photo may receive the label of "a soccer shot" if the user is trying to describe this picture among a collection of general news stock footage for events happened in 2006. Similarly, event history can affect the granularity of event descriptions, e.g., "a lecture on image processing" can be further specified as "lecture 7 of EE480 in Engineering building 332 on image enhancement" if the user has seen the previous sessions of the same lecture series.

4) Sensory Gap: The sensory gap [33], [129] is the gap between a real-world event and the information in a (computational) description derived from a recording or recordings relating to that event. This gap is responsible for the loss of certain event facet semantics after media capture and makes them unrecoverable from the media alone.

Multimedia clips and their annotations typically capture a very small subset of the information relating to the event. For example, the attendees of the technical conference all sample different aspects of the "ACM Multimedia 2006" event. Among the small subset of events one person can participate, they choose to record another reduced set. For example, a user may decide to take photographs at the conference dinner, while recording the speech at the keynote presentation on her mp3 player for later review. Others may choose to document conference presentations on their computer via text. What they choose to record (e.g., a walk on the beach, the conference keynote, dinner) and how they choose to capture is informed by their situational context. Every user who captures the event implicitly leaves out most of the people and the subevents in the conference—e.g., the conference talks not attended, or lunch conversations unarchived. Importantly, while no single user has a complete understanding of the conference event, summarizing and aggregating the capture across users may still recover the semantics of the conference.

The capture may also be a distorted recording to the original experience. For example, the semantics of an event may be altered if the person decides to take only black and white photographs—it is no longer possible to describe the event capture in terms of the color of the clothes worn by the participants. Note that different stages of media production, such as premeditation, capture or editing [56], [96], will change or create new metadata

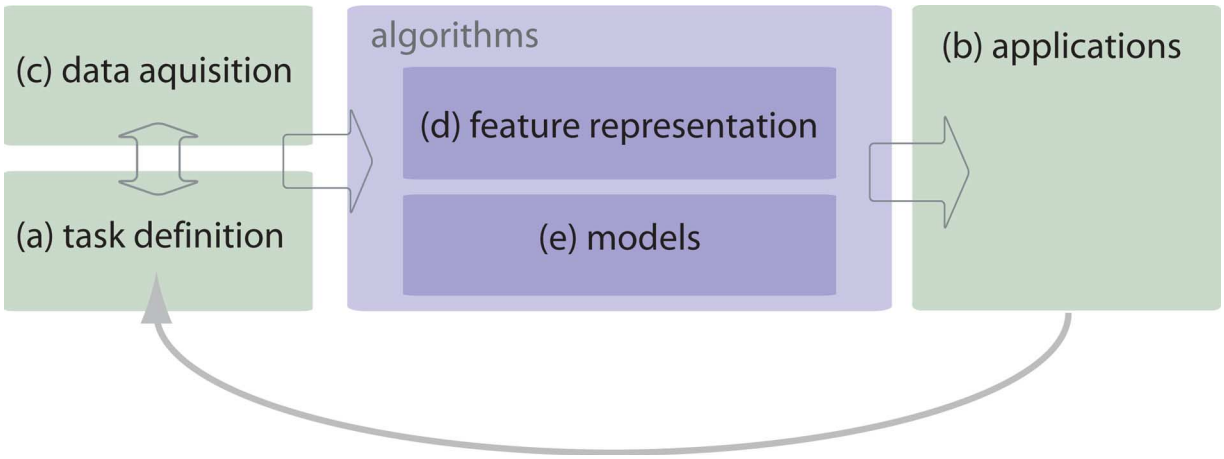

Fig. 5. *General components for event modeling. Components (a)–(e) are discussed in Section III-A through III-E, respectively.*

about the 5W1H in the event being captured. For the purpose of this paper, we do not distinguish these changes incurred by different operations.

III. ELEMENTS OF EVENT MODELS

In this section, we discuss typical event-modeling systems. These systems can be decomposed into a few broad components, as shown in Fig. 5 and Table 1. The main body of this section will be devoted to a high-level overview of each component, presenting their common forms and variations. These components will also help structure the discussions in Section IV by providing a basis for system comparison, as summarized in Table 2.

Fig. 5 depicts the five broad components of an event-modeling system: *task definition*, *data acquisition*, *feature representation*, *modeling*, and *applications*. Intuitively, both the data and analysis operations flow from left to right. An

Table 1 Overview of Event Detection Components. For Descriptions See Corresponding Parts in Section III. Component-Labels and Type-Labels are Used in Table 2

COMP. LABEL	COMPONENT	TYPE LABEL	TYPE DESCRIPTION	SECTION
(A)	event type	(a)	usual, normal	III-A.2
		(b)	unusual, abnormal	
(B)	task	(a)	detection	III-A.1
		(b)	recognition	
		(c)	verification or identification	
		(d)	segmentation	
		(e)	annotation	
		(f)	discovery	
(C)	input	(a)	one continuous capture	III-C.1
		(b)	multiple synchronous continuous captures	III-C.2
		(c)	one edited sequence	III-C.3
		(d)	collections of media objects	III-C.4
(D)	features	(a)	unimodal low-level features: audio/visual/text	III-D.1
		(b)	tracked objects or parts	III-D.2
		(c)	generic mid-level features	III-D.2
		(d)	metadata	III-D.2
		(e)	multiple feature types and modalities	III-D
(E)	models	(a)	distance-based models, nearest neighbors	III-E.1
		(b)	grammar-based model	III-E.1
		(c)	discriminative models: supervised	III-E.2
		(d)	generative models: supervised	III-E.2
		(e)	generative models: semi-supervised and unsupervised	III-E.3
		(f)	other semi-supervised and unsupervised models	III-E.3
		(g)	interactive and semi-automatic system	III-E.3

Table 2 Overview of Event Detection Systems and Their Corresponding Components. For Component Labels See Table 1. For Detailed Discussions See Section IV

references	type	task	input	features	models	data domain	section
[31], [109], [160]	(Aa)	(Ba,Bb)	(Ca)	(Da)	(Ea, Ee)	action sequences	IV-A.1
[53], [61], [65], [71], [92], [126], [141]	(Aa)	(Ba,Bb)	(Ca)	(Db)	(Eb)	surveillance, action sequences	IV-A.2
[19], [52], [102], [121], [132]	(Aa)	(Ba,Bb)	(Ca)	(Db,De)	(Ed)	gesture and action sequences	IV-A.3
[8], [26], [91]	(Aa)	(Ba,Bb)	(Cb)	(Db,Dc,De)	(Ed)	meeting, surveillance	IV-A.3
[40], [43], [150], [154]	(Aa)	(Ba)	(Ca)	(Db,Dc)	(Eb)	surveillance, broadcast news	IV-A.4
[18], [83], [159], [162]	(Aa)	(Ba,Bd)	(Cc)	(Da)	(Ea,Eb)	various broadcast	IV-B.1
[73], [82], [134]	(Aa)	(Ba,Bd)	(Cc)	(De)	(Ea,Ed)	various broadcast, film	IV-B.1
[22], [27], [58], [63]	(Aa)	(Ba,Bd)	(Cc)	(Dc,De)	(Ec,Ed)	broadcast news	IV-B.2
[13], [41], [79], [147]	(Aa)	(Ba,Bd)	(Cc)	(Da)	(Ea,Ed)	sports	IV-B.2
[39], [44], [97], [153], [158]	(Aa)	(Ba)	(Cc)	(Da,Db,De)	(Ea,Ec,Ed)	sports, film	IV-B.2
[142]	(Aa)	(Ba,Bb,Bd)	(Cc)	(De)	(Eg)	film	IV-B.2
[28], [140], [148], [161], [165]	(Aa)	(Bf)	(Ca,Cc)	(Da,De)	(Ee)	meeting, broadcast, surveillance…	IV-C
[42], [46], [54], [55], [104], [108], [164]	(Ab)	(Bf)	(Ca,Cc)	(Da,De)	(Ef)	surveillance, broadcast, lifelog	IV-C
[81]	(Aa)	(Ba)	(Cd)	(Da)	(Ea,Ec)	personal photos	IV-D
[48], [50], [94], [95]	(Aa)	(Ba,Bd,Be)	(Cd)	(Dd)	(Ea,Ee)	personal photos	IV-D
[30], [74], [85], [105], [125], [167]	(Aa)	(Ba,Bd,Be)	(Cd)	(Da,Dd)	(Ea,Ee)	personal photos	IV-D
[128]	(Aa)	(Ba,Bd,Be)	(Cd)	various	various	news, documentary	IV-E
[35], [88]	(Aa)	(Ba)	(Ca,Cb)	various	various	surveillance and UAV	IV-E
[1], [2]	(Aa)	(Ba)	(Cb)	various	various	meeting	IV-E

event detection problem is essentially influenced by the definition and properties of the target event, constrained by the availability of data, and directed by the goals of the intended applications. The data acquisition setup is also influenced by what the system is to accomplish. The acquired data feeds the algorithmic components that extract content features, builds event models, and makes decisions about the events. These additional metadata from features and modeling output then drive multimedia event applications, mainly information-seeking tasks such as retrieval, browsing, or summarization.

Table 1 summarizes a subset of the common options for the five main components described in the rest of this section. The components (a)–(e) in this table also label the algorithms in Section IV and Table 2. Here, the *task definition* is naturally broken into *event type* and system *task*, and event applications are omitted from this table since the generic applications discussed in Section III-B can be driven by most event detection systems, and event-specific retrieval systems (e.g., summarization of medical imagery) are beyond the scope of this survey.

In the rest of this section, we examine various event detection problems currently being addressed (Section III-A), then tie the formulation of event detection problems with event applications (Section III-B) and data acquisition scenarios (Section III-C); we also briefly overview the two major algorithmic elements: feature representations and computational models in Sections III-D and E, respectively.

A. Event-Modeling Problems

Although we often hear about systems that perform *event detection*, the meanings and scopes of what are being solved are very diverse. This diversity comes from two sources. The word *detection* can refer to several different computational tasks, some of which maybe co-existing or overlapping. Furthermore, the properties of an *event* are largely variable at different semantic levels and aggregations along several principle attributes, as described in Section II. Such variation can transform the problem very dramatically. We now examine these two factors in more detail.

1) Different Tasks in Event Mining: The task of associating one or more semantics from data can take one of many forms, depending on the data available and the target decision. Extending from pattern classification problems in closely related areas such as face detection and recognition [59], [163], speaker identification [111], and image segmentation [123] we present the following six tasks on event semantics.

1) Detection—compare data (multimedia clip) with a known event or event model, decide the presence or absence of the event, e.g., "does this shot contain a person walking?"
2) Segmentation—locate which part of the data correspond to the event of interest, this specification can be in time, space, or both, e.g., "when did the game point for Wimbleton men's final start, how long did it last?"
3) Recognition—recover from data a description of the event containing one or more of the five W attributes, e.g., "which word in the American Sign Language does this gesture represent?"
4) Verification or identification—confirm a specific property in the event "is this Agassi's secret serve with a speed up to 160 mph?"
5) Annotation—associating possibly more than one semantic labels to data, possibly choosing from a semantic ontology and taking into account the relationships among the semantics, e.g., "tennis match, crowd, athletes, Wimbleton, semi-final."

6) Discovery—find events without knowing its semantics beforehand, using the regularity or self-similarity among event instances, e.g., from Fig. 4(b), "the crowd gathering at room 332 every Wednesday at 11 am."

Note that these different tasks often co-occur, e.g., detecting an "airplane landing" event and segmenting it out in time. Also note that the first three tasks, i.e., detection, recognition, and verification, would require a known event model or event description, while annotation require more than one event models that may be interrelated. Finally, in current event-modeling systems some of these tasks tend to associate with particular attributes in the "5W1H," e.g., segmentation is typically concerned with *when* and *where*, recognition is concerned with *what* and sometimes *who*, etc. Here, we present the *detection* and *segmentation* as two separate tasks where *detection* does not involve finding where/when the event is; this is consistent with recent benchmark tasks on generic concepts and events [139] and different from the definitions in face detection [59] where the *detection* task encompasses the two.

In the rest of this paper, we use *event detection* in its general sense, which can include one or more of these five tasks above.

2) Event Properties: Regular or Sporadic, Usual or Unusual: Many existing research in event detection carry one or more of the common modifiers for their detection target, for example, "sporadic events," "rare or unusual events," "recurrent events." These modifiers constrain event detection problems differently and they exercise important influences on the algorithm being chosen.

Being *regular* means "marked or distinguished by steadiness or uniformity of action, procedure, or occurrence" [6], while being *sporadic* is "characterized by occasional or isolated occurrence, appearance, or manifestation" [6]. We can aggregate regular event instances into an event class as described in Section II-B1 by the steadiness or uniformity of their occurrences, e.g., "department seminars Fridays at 11:00 am," "the serves and returns in a tennis game" (with every play started with a serve and followed by one or more returns). While *sporadic* events may have consistent meaning but usually cannot be explained or predicted from the regular events, pertaining to the *why* aspect of the 5W1H, e.g., "aces in a tennis game" (serves that the opponent cannot return), we know that a tennis game must contain a number of plays for a single point, which in turn contain serves and returns, but it is impossible to predict whether an ace will happen based on the plays happened so far.

An alternative classification of event types is *usual* versus *unusual*, or *normal* versus *abnormal*. Being *unusual* or *abnormal* not only means that the *why* aspect is unaccounted for, but also means that the *what* and *how* aspects are unknown from a normal pool of media clips. For example, a "foul in a soccer game" is a *sporadic* but normal event, while a "banana kick" or an "upheaval of soccer fans" are rather *unusual* in collections of soccer broadcasts.

These event properties affect our algorithm design choices. For example, *regular* and *sporadic* events in sports can both be detected with a set of trained classifiers [39]; detecting *unusual* events often means finding outliers that do not fit the current set of models [164]; once "discovered," we can also build models to describe the *unusual* events [161].

3) Current Event Detection Problems: Among the event detection problems currently being addressed, we can see a few salient groups with very similar problem setups. The similarities are in the semantic aspects or along event-modeling components: detecting which of the five W attributes, how to group or aggregate event instances, properties of target events, data format, and target applications. Section IV reviews existing research work in the following four groups in more detail.

1) detecting known events and activities from one continuous capture (Section IV-A);
2) event detection in edited sequences (Section IV-B);
3) unsupervised event detection and pattern mining, i.e., detect *unknown events* (Section IV-C);
4) event annotation in a collection of media objects (Section IV-D).

We simply list the partitions here and leave their problem definitions, scopes, examples and discussion to the individual subsections in Section IV. The main purpose of this partition is to facilitate the presentation of a diverse collection of existing work, rather than drawing artificial boundaries on what is worth working on. Aside from shared problem setup and goals, the emergence of these partitions can in addition be attributed to the perceived usefulness along the criteria for maximizing the impact of content-based modeling and metadata extraction [24], which include providing metadata that are neither available from production nor easily generated by humans and working on content collections that will most benefit from the value-add—those of large volume and low individual value.

B. Applications of Event Modeling and Detection

In order to understand what an event detection system should achieve, it helps to examine the uses of event metadata. Event metadata can provide semantically meaningful indexes that help decompose a task with faceted metadata and map an information-seeking task to a multimedia event ontology, such as the large-scale multimedia ontology (LSCOM) for broadcast news [98]. This can provide additional aspects for matching and filtering information just as the faceted attributes of author, title, publisher, helps catalog, search, and promote books in libraries.

Many information-seeking tasks, on the other hand, can be mapped onto such event-based semantic metadata. For

instance: 1) Active, or goal-oriented information seeking involves finding media clips that matches an existing description, the 5W1H can help narrow down the range of possible media clips. 2) Matching involves deciding whether or not a media clip matches a given description, e.g., "is this video a machine learning lecture?" "is this video funny?"—for which semantic metadata can be directly checked for the match. 3) Browsing and impression formation means trying to get an idea about the content of an entire collection from an overview or random sample of the content, e.g., the "most popular tags" page at Flickr [5], or the thumbnails view in modern operating systems. Semantic metadata can be directly summarized into text form, tag clouds, or thumbnails, which are more intuitive and convenience for an overview than collections of media sequences unrolled in time. 4) Indexing and archival is to insert metadata for items in a collection so that they can be easily found at a later time, e.g., a librarian inserting library-of-congress call numbers for newly acquired books. Here, event metadata are directly applicable as additional indexes.

The benefit of semantic metadata on information-seeking tasks can propagate its influence to real-world multimedia systems, for example, semantic metadata has helped generate significantly better results for automatic and interactive video search [80], [99], [100], [130].

C. Forms of Event Media Data

Multimedia archives are snapshots of real-world events from capturing, editing, and archiving with limited metadata and annotation. The setup for media capture imposes limits on what are available to us for data analysis, it also puts constraints for system design. Here, we examine a few typical scenarios and examine the often implicit but important assumptions.

1) Single Stream From One Continuous Take: This scenario typically uses one camera and/or audio recorder, with either a fixed installation such as close-circuit surveillance [133] or moving in space such as unmanned aerial vehicle (UAV) [88] or lifelogs [28], [42], [133]. The scope of data analysis is within a start/stop of the recording device. Thus, there is continuity in both space and time. Mapped to the 5W1H in event attributes (Section II-A), the offset between the media time and the real-world time is constant [as shown in Fig. 2(a)], and the location correspondence is either fixed or continuously changing. Such continuity enables common signal-level processing operations in the image sequences and sound such as motion estimation, tracking, registration, background substraction, and mosaicing. This scenario is addressed by many visual-based event and activity modeling systems (Section IV-A) due to its high value in practice (security systems and other monitoring services) and the simple fact that the other scenarios are various combinations and compositions of the start/stop of recording.

2) Multiple Concurrent Streams: Multiple cameras or microphones can be set up to capture multiple view points. This setup is commonly found in surveillance networks [88] and meeting recordings [133]. These concurrent streams offer richer representation about the original scene. They are often calibrated so that the spatial and temporal correspondence can be reconstructed, yet they present additional challenges in data association for finding events from multiple sources. Existing work on analyzing multiple concurrent streams is reviewed in Sections IV-A and C.

3) Single Stream From Multiple Takes: Conventional video and audio are linear media in that they contain a single sequence meant to be consumed in temporal order. Such a sequence can be obtained by concatenating segments taken at different time and/or location, as shown in Fig. 2(b)–(d). Most broadcast content and its raw footage are in this category, e.g., the TRECVID corpora containing news, documentary, and rushes [139]. Event semantics in these streams manifest themselves not only within each shot but also in the syntactical relationship in a sequence of shots or a sequence of scenes [124]. Shot boundaries introduce discontinuities in time and space in these streams and the reference times and locations of such discontinuities are often unknown. This unknown correspondence and the typically short shot length (a few seconds) can prevent low-level vision algorithms, such as tracking and object segmentation, from performing robustly. Existing work on analyzing events in these edited sequences is reviewed in Section IV-B.

4) Media Collectives: Real-world events are also often captured in collections of loosely related media streams and objects, such as pictures from vacation trips [94], user-generated content around breaking news [7], or photo pools on community events [4]. Each media object in the collection may be an unedited continuous capture or a produced stream as described in Sections III-C1 and C3, respectively. Yet they tend to be temporally asynchronous and spatially dispersed with unknown correspondences. These collectives provide comprehensive views of the events of interest, yet the appearances of different media clips are usually diverse. Therefore, media collectives challenge algorithms that rely on audio–visual, or spatial-temporal continuities. Existing research on analyzing a collection of media objects is reviewed in Section IV-D.

D. Feature Representation

Feature representations are extracted from media sequences or collections, converting them into numerical or symbolic form. Such representations are convenient system representations and are prerequisites to event recognition. Good features are able to capture the perceptual saliency within the event, distinguishing it from other events, as well as being computationally and

representationally economical to lower recognition cost and improve performance. It is beyond the scope of this paper to provide a comprehensive survey of audio–visual features. We present a summary of commonly used features for completeness and direct the users to respective surveys on image, video, speech, and audio features [23], [49], [72], [129].

In order to structure the discussion, we group the features across different media types into three common categories, based on methods for computing them and their level of abstraction.

1) Low-Level Features: Low-level features directly reflect the perceptual saliency of media signals. The procedures for computing them do not change with respect to the data collection or the event being detected.

Still images are usually described in three perceptual categories, i.e., color, texture, and shape [129], while image sequences introduce one more dimension of perceptual saliency, i.e., motion. Color features are popular due to their ability to maintain strong cues to human perception with relatively less computational overhead. The main concern in reliably extracting color information is to choose from a variety of color spaces and achieve perceptual resemblance and color constancy over different scene and imaging conditions. Local shapes capture conspicuous geometric properties in an image; this is among the most-studied image features, since psycho-visual studies have showed that the human visual system performs the equivalence of edge detection [62]. Local shapes are often computed over local gray-scale or color derivatives. Texture loosely describes an image aside from color and local shape. It typically reflects structure and randomness over a homogeneous part of an image. Filter families and statistical models such as Gabor filters and Markov analysis are popular choices for capturing texture. Motion provides information about short-term evolution in video. The 2-D motion field can be estimated from image sequences by local appearance matching with global constraints, and motion can be represented in various forms of kinetic energy, such as magnitude histogram, optical flows, and motion patterns in specific directions. Although color, shape, texture, and motion can be described separately, there are features that provide integrated views such as correlogram [64] (color and texture) or wavelets (texture and local shape).

General audio can be characterized by a number of perceptual dimensions such as loudness, pitch, timbre. Loudness can be captured by the signal energy or energy in different frequency bands. Primitive pitch detection for monophonic tonal signals can be done with simply counting the zero-crossing rate. More realistic pitch detection involves autocorrelations and various modifications. Timbre is captured by the amplitude envelop of spectrograms as well as the dynamics of the sound, i.e., the relative strength of different harmonics for tonal sounds and their onsets and attacks. More elaborate features for modeling each of these aspects exist, such as robust pitch extractors [34], [89], linear prediction coefficients (LPC) [107], and frequency-warped spectral envelops such as the mel-frequency cepstral coefficient (MFCC) [49].

For text annotations or for audio signals that contain speech, features can be computed on the text or speech transcript, using simple measures such as word counts. Compared to text annotations only, speech signals have additional timing information upon which prosody features such as speaking rate and pause length can also be computed.

2) Mid-Level Features and Detectors: Mid-level features are computed using the raw signal and/or low-level features. Their computation usually involve signal- or data-domain-dependent decisions in order to cope with the change in the data domain and target semantics, and sometimes training is needed.

Mid-level features capture perceptual intuitions as well as higher level semantics derived from signal-level saliency. Examples of mid-level features and detectors include: tracked objects and segmented object parts [157]; visual concepts pertaining objects, scenes and actions, such as people, airplane, greenery [139]; audio types, such as male/female speech, music, noise, mixture [120]; and named entities extracted from text passages [29]. There are also mid-level features that are specific to a data domain, such as the crowd cheering detector or goal post detectors in sports videos [39].

Features cannot only be extracted from media content, they can also come from the 5W1H in faceted metadata, i.e., structured attributes fields such as dates, location proximities [94], semantic distances between locations [125], etc.

3) Feature Aggregates for Recognition: Feature aggregates are derived from features and detectors so that the inherent spatial-temporal structure in the media sequence can be represented as numbers/vectors/bags so as to fit the data structure required by most statistical pattern recognition models. In practice, this aggregation is usually done with one or several of the following operations.

1) Accumulative statistics. This includes histogram [135], moments [137], and other statistics over collections of points. These statistics provide simple yet effective means for aggregating features over space and time. They have the advantages of being insensitive to small local changes in the content as well as being invariant to coordinate shift, signal scaling, and other common transformations. The associated disadvantage is in the loss of sequential or spatial information.
2) Point selection. The selection of possible feature detectors from candidate portions of the original signal aims to preserve perceptual saliency and provide better localization of important parts. Tracking and background substraction can be

viewed as one type of selection, as well as salient parts extraction and [86], silence detection in audio, or stop word removal.

3) Set aggregation. This is done over the features in an image, a image sequence, audio segment, or other natural data units. The sets can be unordered or ordered, e.g., bag of words, bag-of-features, sequences, or more general graph structures.

4) Discussion About Features: The separation we made among low-level, mid-level, and feature aggregations is sometimes blurred. For example, tracking can be seen as either a mid-level feature extraction or part of the selection process. Also, note that selection and aggregation can also happen before the extraction of features, such as silence removal, stop word removal, etc. While good features are deemed important, some prefer a featureless approach [77] that leaves the task of determining the relative importance of input dimensions to the learner. With the wide variety in feature representations, choices shall be made from domain knowledge and the event modeling task at hand, and coming up with the "optimal features" would remain an open problem.

E. Computational Models

In event detection, models are responsible for mapping data representations to semantic descriptions, where the descriptions are either in the forms of a discrete label (e.g., person running) or continuous states (e.g., the pose or velocity of an object). The richness of computational models warrants an entire book [57], and we refer the readers to existing texts and reviews for pattern recognition and machine learning approaches [57], [67] for a comprehensive treatment. We use this subsection to present a few observations on choosing and using models for event detection.

1) Knowledge-Driven and Data-Driven Approaches: Human perception of sensory streams are known to be both knowledge-driven and data-driven [90]. Several well-known event recognition systems from the 1990's are mainly knowledge driven, using automaton [92], finite state machine, or grammar models for inference. Data-driven models range from variants of nearest neighbors to the generative and discriminative statistical models that represent complex class boundaries and encode relationships among the input and output. Nearest neighbor, or distance-based classifiers remember the primitives of known classes and classifies and then classify new examples to the nearest primitive; this has been extensively used in many applications, such as face recognition [15] or action recognition [31].

It was observed in speech recognition research [70] that large amounts of annotated training data would enable data-driven approaches to outperform its knowledge-driven counterpart. Similar phenomena is also observed with increasingly large amount of visual- and multimodal event repository being collected and made available to the research community (Section IV-E). One example that combines knowledge and data is stochastic context-free grammar (SCFG) [65]. SCFG is initialized and weighted by data, it smooths the HMM equivalent with nontrivial weights among unlikely or unseen parse strings and does not suffer from the lack of data to reliably estimate or even foresee unlikely paths. Ivanov and Bobick [65] found that SCFG outperform hidden Markov models (HMM) in the human activity recognition task due to the inability of HMMs to represent a large variety of possible paths. Having weighed the pros and cons, smart combinations of knowledge and data or systematic ways to encode knowledge into data-driven models are very much desirable.

2) Generative and Discriminant Models: Generative models "produce a probability density model over all variables in a system and manipulate it to compute classification and regression functions" [68]. Discriminative models directly attempt to "compute the input–output mappings for classification and regression," eschewing the modeling of the underlying distributions.

Discriminative methods—logistic regression, support vector machines (SVM), boosting—have seen strong success in both research and practice over the last few years. Generative models—HMM, dynamic Bayesian network (DBN), linear dynamic systems—are still the models that many choose for capturing events that unfold in time (Sections IV-A–C). The popularity of generative models is due to two reasons: they offer to "explain" the data in addition to being able to complete the detection task, and they are naturally suited to capture the structure of the data (sequence, relations). These models with structural constraints do not suffer a search space of K^L, with K the number of possible states and L the length of the sequence. Discriminative models with specifically designed feature representation (e.g., bag of features [113], fisher scores [66]) and a similarity metric (e.g., Earth-Mover's Distance [116], string kernels [84]) have also shown good detection performance in domains like computational biology and text classification. Discriminative models have also been used to model video events such as story segmentation [63] or short-term events [40], [150], [154] with promising results.

3) Continuum of Supervised, Unsupervised, and Semi-Supervised Models: A general machine learning task involves learning a mapping from input space X to output space $Y : f(X) \rightarrow Y$. For *supervised learning*, Y is known at training time, while in *unsupervised learning*, Y is unknown. In supervised learning, only $f(X)$ is learned, in unsupervised learning $f(X)$ and Y are estimated at the same time, while in semi-supervised learning a subset of Y may need to be learned together with $f(X)$, or Y may need to be learned with certain constraints (equivalence, mutual exclusion, sequential, etc.). Semi-supervised learning

methods "use unlabeled data to either modify or reprioritize hypotheses obtained from labeled data alone" [166]. Popular semi-supervised approaches include EM with generative mixture models, multiple instance learning, self-training, co-training, transductive support vector machines, and graph-based methods. We would direct the readers to a separate survey [166] for details of these approaches.

Event detection and recognition problems do not readily map to this classic supervised versus unsupervised setup, since event data are inherently structured, and both X and Y can come in different granularities. For example, a data tuple (x, y) can mean any of the following: a pixel x has label y, a region x has label y, an image x is assigned label y, at least one image x in a video sequence is assigned label y, or the entire sequence x share the label y. Therefore, the distinction of supervised and unsupervised models for event recognition is a gradually changing grayscale, rather than being black or white. The level of supervision varies depending on what kind of labeling information is available at training time. This information can include: a sequence-level label about whether an entire clip contain an event, its start/stop time, the object bounding box or parts, if it is possible or not for two events to co-occur, etc. These diverse scenarios makes various semi-supervised learning algorithms very desirable. When formulating event detection as a learning problem, deciding what to label (pixels, frame, or sequence) and how the data should look can be more important than building the machinery to learn the mapping from data to label.

F. Discussion

It is worth noting that the five components in Fig. 5 are not necessarily separate. Different types of feature extraction process can be interwoven (Section III-D4), as can feature extraction and modeling, or data capture and feature extraction. For instance, learning similarity measures or learning to select features resides in the intersection of features and modeling. Data capture and feature extraction may be done in one shot, with embedded architecture such as smart camera systems [146]. Implemented in hardware, such design not only improves upon software-based system on detection speed, it also makes the deployment of event detection systems easy for both everyday use and large-scale multimedia sensor networks.

IV. EVENT-MODELING SYSTEMS AND EVALUATIONS

Having discussed the problem space of event modeling and the general components of its solutions, we now turn our attention to describe a few commonly addressed scenarios in the literature. Table 2 contains a roadmap for this section, anchored by the different system components described in Section II as well as the data domains they are applied to.

A. Detecting Known Events From One or More Continuous Capture

Audio–visual streams from one continuous capture is a frequently studied data domain for event detection. This type of data is often found in many real-world applications such as closed-circuit surveillance, UAV video, and video input for human computer interaction. Moreover, it is the building block of edited sequences and larger media collections.

The dual continuity of space and time allows events to be detected as "long-term temporal objects" [160] by analyzing the behaviors of one or more foreground objects in a static background. Feature extraction in one continuous capture or multiple synchronous captures typically involve differentiation in space and/or time with pixel intensities, segmentation of foreground/background, or extraction of moving regions, objects or parts. This is done with a variety of techniques such as image stabilization, registration and mosaicing, background substraction, and object and region tracking. For coverage on the extensive literature on foreground extraction and tracking, we refer the readers to existing surveys [21], [157]. In the rest of this section, we review some examples of event recognition from 2-D images, grouped by their inference mechanisms.

1) Distance-Based Action Recognition: Human actions can be inferred by comparing distances of pixel-based features that represent changes in space and time. This approach is effective because distances of pixel-based features are well defined, and such recognition works well on a constrained domain. Davis and Bobick [31] use a two-step approach for movement recognition in smart rooms. The first step constructs a binary *motion energy image* indicating the presence of motion on the 2-D image plane, the second step computes a *motion history image* by integrating the motion intensity image weighted by the recency of motion. Action recognition is then achieved by computing the vector-space distance with the template motion history images. Zelnik-Manor and Irani [160] detect events as a multiscale temporal aggregate within a continuous video shot. They use the directions of local intensity gradient as the feature representation of input images and then use χ^2 divergence measure to cluster the feature distributions at multiple scales. Rao *et al.* [109] use the motion curvature feature of the object (hands) and perform dynamic segmentation of the object in both space and time, producing a mid-level primitive called "instants." The time, location, and sign of the curvature are then matched to templates for event recognition. The first two approaches above are invariant to the execution rate of the action, and the third one is somewhat invariant to view point changes. Time-scale invariance is achieved with the design of a distance metric over accumulated feature statistics over the entire sequence, while view-invariance results from the design of a robust feature, i.e., the sign of motion curvature.

2) Grammar-Based Approaches on Tracked Object Parts: Events can be viewed as structured aggregation and evolution of tracked objects and their parts. Grammars and graphs are natural choices to encode these relationships. Medioni *et al.* [92] represent objects as a graph, with tracked object parts as the nodes and the tracking likelihoods as edge weights. The object trajectories are then combined with a "location context" including static objects in the scene as well as other moving objects. Events are matched with known event classes using finite state automaton. Hongeng *et al.* [60], [61] parse tracked objects into individual action threads, the action threads are then matched to event classes using variants of stochastic finite automaton: binary interval-to-interval networks or temporal logic network for multi-agent events. Ivanov and Bobick [65] use SCFG to recognize complex action sequences (e.g., conducting music) from video. This is achieved by first segmenting and tracking low level primitives (e.g., hands or vehicles) using statistical detection and then use SCFG to parse the sequence of primitives taking into account substitution, insertion, and deletion errors. Although it relies on the immediately preceding predicate, SCFG is deemed superior than HMM here if the sequence evolution is complex and if not enough training data are available to reliably estimate the probabilities. Shi *et al.* [126] propose propagation networks to recognize activity as partially ordered part sequences. This partial order is represented by a set of constraints including the duration, temporal precedence, and logical constraints with temporally coexisting or adjacent parts, and the inference process with a propagation network is done with particle filtering. Compared to SCFG, propagation networks do not need to explicitly represent all valid event orders in the grammar. Joo and Chellappa [71] extend SCFG with attribute grammar, which can specify feature constraints on the part symbols and contains and instance of SCFG in the model. This was used to detect normal and abnormal events in parking lot videos. Hakeem *et al.* [53] adopt a hierarchical representation of events that consist of subevents and case lists, which is solved with subtree isomorphism at detection time.

Grammar-based matching algorithms call for formal language structures to represent relationships and constraints. The Video Event Representation Language (VERL) [47], [101] is an introduction to an ontology framework for representing video events. VERL also has a companion annotation framework, called Video Event Markup Language (VEML). This work makes the description of events composable, whereby complex events are constructed from simpler events, e.g., agents and their actions, by operations such as sequencing, iteration, and alternation. Velipasalar *et al.* [141] use a similar event definition language that encodes logical relationships in motion descriptors and objects tracked in a multicamera surveillance network to recognize events such as "tailgating" and "person walks by elevator and then exits building."

3) Generative Models for Event/Action Inference: Given enough training data, statistical models are often preferred for learning on structured input/output in many applications, such as in speech recognition [70]. HMMs, being a popular choice for stochastic representation of sequences, are used by Schlenzig *et al.* [121] to recognize four types of gestures and by Starner *et al.* [132] to recognize American Sign language from wearable computers.

Extensions of HMM have been developed to account for multiple sequences, multilayer dependencies, and other complex data structures. Coupled HMM (CHMM) [19], [102] explicitly models the temporal dependencies among different streams, such as audio, video, user input for multi-object multi-agent action recognition. Chen *et al.* [26] use dynamic Bayesian network (DBN) to detect social interaction from multicamera nursing home surveillance videos in a two-level setup. The first level is the processing of audio–visual streams to locate segments with any human activity and track their 3-D coordinates, using moving regions identified with background substraction, fused with energy-based audio features. The second level uses DBN for inferring events such as "walking assistance" and "standing conversation." Zhang *et al.* [91] analyze multicamera/microphone meeting captures for group interaction events such as discussion, monologue, presentation, and note taking. The audio and visual streams are processed independently to generate a set of features relating to skin-colored blobs, audio localization, pitch, and speaking rate. A two-layer HMM is then used to infer individual action and group action in cascade. Each state in the HMM is assigned domain-specific meanings and the parameters are learned from data. Aghajan and Wu [8] also use multilayer graphs to infer gestures from multiple cameras, where the lower layer performs "opportunistic fusion" of simple features within a single camera and the upper layer takes care of "active collaboration" between cameras. Gupta and Davis [52] unify object recognition and tracking and event recognition, aiming to disambiguate objects with temporal context. They also choose Bayesian networks for modeling interactions between human and objects, where the nodes in this belief network correspond to either the object or the types of motion such as reach, manipulate, etc.

4) Discriminative Models for Event Detection: While graphical models are natural choices for modeling temporal evolution in one continuous stream, discriminative models have also shown good performance since they directly optimize for the detection boundary. Eng *et al.* [43] detect drowning and distress events in swimming pool surveillance videos. A variant of neural networks called functional link network is used on extracted foreground objects (people), compensating for the aquatic background environment. Kernel-based classifiers have been used to detect events from within a shot broadcast content. Shots in TV broadcasts tend to be only a few seconds long. The

variations in scene, lighting, and camera conditions are typically too large to do reliable background substraction or tracking, let alone seeing consistent object appearance and trajectories. In this scenario, non-object-specific visual features and mid-level concept detectors become useful for distinguishing a shot that contains a generic event (e.g., *airplane landing*, *riot*) from those that do not. SVMs can be used on kernels generated from HMM likelihoods and parameters from input feature streams [40]. Kernels can also come from bag-of-features representation of temporal streams, with the similarity metric computed with earth-mover's distance (EMD) or multiresolution temporal match [154]. Multiple available kernels can also be combined [150] to learn both the class decision and combination weights simultaneously.

5) Discussions: Event recognition from one continuous capture, as reviewed in this section, focuses on recognizing the *what* attribute in the 5W1H of event descriptions (Section II), the *when* and *where* attributes are continuous or assumed to be fixed.

The reconstructions of 3-D plans for events or actions have been explored [103], [136]. We have mostly reviewed event detection from 2-D image sequences. The 2-D representations in image sequences introduce an inherent limitation of view dependence, while the 3-D approaches suffer from larger search space and an ill-formed reconstruction problem.

Most of the work mentioned here is based on visual information. A few studies also included audio features [25], [26], [91] and found significant advantage in doing so. Multimodal multisource event detection is likely to receive more attention in the near future due to many emerging applications and the availability of large multimodal collections.

While most of this section covers the detection of *known* events, Section IV-C discusses the detection of *unknown* events based on self-similarity and regularity, many of them also on one continuous media capture [140], [161], [164], [165].

B. Event Detection in Edited Sequences

The meanings of a continuous, edited multimedia sequence reside both in each shot and in the syntactic relationships among adjacent shots created by the director. The most prevalent forms of such content are feature films and television broadcast archives, where generic and domain-specific events are useful for indexing, search, and summarization. A broad definition of events in produced videos include two categories: those resulting from video production (i.e., camera or editing operations), such as shot boundaries, scene change; and those are inherent in the video content, such as changes in objects, settings, or topics.

1) Detecting Production Events: Detection of video production effects is a natural first step towards breaking down the video understanding problem, and it has received considerable attention since the beginning of multimedia analysis [162]. Shot boundary detection is typically detected as a change in color, texture, or motion features [18]. Recent benchmarks show [139] that abrupt and gradual changes in broadcast content can be detected with ~90% accuracy, and statistical models such as SVM [83] and graph transition models [159] have shown good performance. Several shots with consistent location and ambient sound constitutes a scene. Scene changes in films can be inferred [73], [82], [134] with features related to chromacity, lighting, audio features, coherence, and memory models.

2) Detecting Content Semantics in Produced Videos: Domain-specific events inherent in video content can be further categorized into regular patterns or spontaneous events. Some video domains contain recurrent continuous semantic units with a coherent, comprehensible meaning, such as stories in news, and "plays" in many kinds of sports. Such units are common and the detection of them is likely to rely on features that reflect the content and the production conventions. A news story is "a segment of a news broadcast with a coherent news focus which contains at least two independent, declarative clauses" [139]. State-of-the-art detection algorithms achieve good segmentation results, with an F1 measure up to 0.74 [22], [27], [58], [63]. This is done by employing machine learning techniques such as SVM and HMM, along with judicious use of multimodal features such as shot length (production effect) or prosody in the anchor speech (content feature).

In sports videos, *play* is a common class of basic semantic applicable to many sports, including soccer, baseball, American football, sumo wrestling, tennis, badminton, and so on. Plays can usually be distinguished from the visual information, especially type of shot and camera motion, since broadcast sporting events typically take place with similar scenes and visual layout. Therefore, many play detection algorithms [13], [41], [79], [147] use color, motion, court layout, and tracking features followed by either rule-based or statistical models such as Bayesian network or HMM. Spontaneous events in sports are intuitively characterized by distinct audio cues such as audience cheering and excited commentator speech, particular view angles such as the soccer goal post and penalty area, behavior of salient objects such as players and balls, as well as mid-level detectors such as whistle and goal posts [39], [44], [117], [153], [158]. Models for inferring sporadic events include rules and distances [153], SVMs [39], [117], [158].

Detection of spontaneous events from films or TV drama, such as explosion, clapping, and waterfalls has also been done similarly by Naphade *et al.* [97] with global audio and video features and a probabilistic factor graph, or in a interactive frame work that helps the user to label a

subset of the 5W1H attributes [142]. For generic produced videos, little can be assumed regarding the camera, scene, or objects, due to the frequent transition among shots and the large variations in the scene and imaging conditions. Therefore, the analysis systems typically resort to global content features or generic mid-level features such as color histogram, correlogram, visual, and audio classes.

Most systems focus on the detection of the *what* facet among the 5W1H of events, whereas *who* and *when* are implicit: the director makes a scene cut, the (tracked) player scores a goal, or an explosion happens ten minutes into the film. Explicit labeling for *who* [12], [16], [119] or *where* [156] exploit the correspondences between the visual information and the spoken content.

C. Unsupervised Event Discovery

Most of the work in the previous sections detects *known* events. Automatically detecting *unknown* events can also be very useful when the user needs to explore a new collection, find new things that are unaccounted for among the set of *known* events, or initialize models and data annotations for more accurate modeling. This scenario has received considerable recent attention, in part because media collections have outgrown the amount of reliable annotation. Such videos are available on the web [14], in benchmark activities [139], and from individual research projects such as the Human SpeechHome [115]. There is a large variety of problems and solutions in this area, although it is relatively new. The problems are in two categories, namely, finding *regular* events and/or *unusual events*. The computational models typically involve clustering algorithms, association and co-occurrence mining, and dynamic graphical models, in combination with outlier-identification and model adaptation. The data domains span many raw and produced content types, such as broadcast news, sports, surveillance, lifelog, meetings, etc.

Regular patterns are typically found using clustering operations with various features and models. A continuous media sequence can be either presegmented into fixed-length units or jointly clustered and segmented by generative models typically in the HMM/DBN family. Clarkson and Pentland [28] cluster ambulatory audio–visual streams with HMMs to identify different user locations. Xie *et al.* [148] find that recurrent frames and shot sequences in sports and news programs often correspond to domain-specific multilevel motifs found using hierarchical HMM. Ellis and Lee [42] cluster and segment wearable audio device recordings into homogeneous "episodes" corresponding to locations and activities, using acoustic features over long time windows and spectral clustering. Turaga *et al.* [140] build linear dynamic systems (LDS) on optical flow features for surveillance action events, with the system iterates between model learning and sequence segmentation. The authors also built temporal, affine, and view invariance into the model with the nature of LDS and modified distance metric. Feng *et al.* [44] use association rules on mid-level audio–visual features to discover events in basketball games.

Unusual events are often defined relative to *usual* events with an underlying distance/similarity metric, where *unusual-ness* is captured as deviation from the usual collection with measures such as a very low data likelihood. Zhou and Kimber [165] model multiple surveillance streams as coupled HMM, trained with the usual events and detects unusual events as outliers in the likelihood values. Zhang *et al.* [161] learn unusual event models from audio–visual sequences by adapting from a general usual model. Radhakrishnan *et al.* [108] treat video segments that deviate from the majority of spectral clusters as outlier events and thus find highlights from broadcast sports or surveillance videos. Zhong *et al.* [164] analyze the co-occurrence matrix among video segments and use matrix co-embedding to identify outliers from a variety of surveillance videos with diverse event types. Petrushin [104] detects frequent and rare events in a multicamera surveillance network using multilevel self-organizing map (SOM) clustering on foreground pixel distribution in color and spatial location; visualization and event browsing are anchored with a "summary frame" accumulating all foreground pixels. Hamid *et al.* [54], [55] use n-grams and suffix trees to mine movement patterns seen from ceiling-mounted cameras, finding usual events such as "fedex delivery" and unusual ones such as "truck driving away with its back door open."

Simply mining events based on recurrence or unusual-ness is a first step towards interpreting the results and making them useful. One way to help interpretation is to associate clusters and patterns with words, such as temporal clusters from news videos [149] or motifs from sports videos [45], using models built on the co-occurrence statistics between clusters or motifs and words. One step towards using and further refining clusters is to build supervised classifiers based on the mined clusters. Fleischman *et al.* [46] find frequent motifs from video streams and build SVM classifiers to distinguish household events such as "make coffee," "wash dishes," based on those motifs. Xing *et al.* [152] use undirected graphical models to find cross-modal hidden structures in news videos and use the results to improve concept detection.

While event discovery usually relies on *unsupervised* approaches, the separation between *supervised* and *unsupervised* is, in fact, gray-scale. This is especially true since the inputs and outputs for discovering multimedia events are structured, instead of being i.i.d. as in classic machine learning settings. Here, *supervision* can mean knowing the onset/offset of events but not knowing the location and action of objects or knowing objects and scenes but not knowing their actions and interactions, while being *unsupervised* will still require domain knowledge and clever feature engineering to steer the discovery towards meaningful directions.

D. Events in Media Collectives

The sections above mainly focus on events within one continuous stream or a set of synchronous streams. In a wide range of real-world scenarios, however, image and video are captured asynchronously in time and space, often by different people from various perspectives and sometimes accompanied by textual descriptions. Such scenarios include photo journalism, consumer photo and videos collections, as well as *user generated content* [7] for news and surveillance. In these domains, content streams can be viewed as a collection of media objects, each of which can be described by the 5W1H attributes while the collection reflect an aggregate event as a whole (e.g., Fig. 3).

We can analyze such collections to infer various aspects about their semantics. Consumer photo streams have received much attention as the now ubiquitous digital cameras have made managing ones' own picture collection a challenge. Much work has gone into understanding how people construct, manage, and use their photo collections. Multiple studies [50], [76], [94], [112] show that users group photos by real-world events, and photo collections are recalled by time, location, or the rest of the 5W1H. Time information attracts the most attention among the 5W1H attributes, as it often is an unambiguous event indicator within a small social circle (e.g., family outing last Sunday), and it is available through the EXIF [69] metadata. Using the capture time of photos alone can help segment photo collections into events. For instance, Graham *et al.* [50] obtain segments with locally adaptive threshold, while Gargi [48] finds the larger intervals between typical "bursts" of photos. Content features are also used to help time-based segmentation. Platt *et al.* [105] use color features to further segment large temporal clusters and compensate for corrupted image capture times. Loui and Savakis [85] use block-based histogram correlation in an iterative clustering process involving time and content features. Cooper *et al.* [30] incorporate content feature with temporal information into to a multiscale clustering and segmentation process. Lim *et al.* [81] classified events, objects, and locations using their semantic relations from a predefined photo taxonomy.

Important as time seems for representing events, recent studies have also focused on inferring events from other aspects of the 5W1H in media collections. Resolving a singular event's semantics may appear to be too challenging, yet the main reason why these research efforts have shown some success is in exploiting the inherent correlations among the 5W1H across the captured media. Intuitively, this can include spatial-temporal correlation (a person can only be in one location at any given time) or the social activities of the people involved in the capture (correlations among *who* and *what*). For the spatial-temporal correlation, Naaman *et al.* [94], [95] capture long-term user activities with location and time information, in order to annotate people from temporal, location, and co-occurrence in events and individual photos. For people-event correlation, collaborative annotation systems [125], [167] take into account semantic similarity and co-occurrence and trust to recommend tags for each user based on a faceted model with the 5W1H and the image content.

E. Benchmark and Evaluations

The flourishing collection of object and event detection methods calls for evaluation on larger datasets beyond a few sequences collected in the lab. The purpose is to draw sensible conclusions about different approaches by eliminating the large variance introduced by different datasets. Such evaluation also calls for algorithm and system comparisons that ensure scalability in computational complexity. Information retrieval (IR) benchmark campaigns trace back to 1991, since the inception of Text Retrieval Conferences (TREC) [138], motivated by the realization that IR tasks need to scale up in order to be realistic. These benchmarks are based on a shared dataset, target task, and evaluation metric. They are attractive to researchers as an open, metric-based venue to validate ideas. Moreover, they often also foster collaboration and the sharing of resources. Recognizing the importance of event detection and modeling problems in multimodal interaction and surveillance, a few international benchmark series have been underway in recent years. Existing benchmark campaigns have been mainly on TV broadcast, surveillance, and meeting recordings.

The TRECV Video Retrieval Evaluation (TRECVID) [128], [139] is an international benchmark campaign run by NIST (U.S. National Institute of Standards and Technology) that first started as a subtask in TREC and grown out as an independent activity since 2003. TRECVID has included a number of data genres around TV broadcast archives: vintage documentaries (2001–2002), U.S. News (2003–2004), multilingual news (2005–2006), documentary (2007), and preproduction footages from BBC (2005–2006). The benchmark tasks include shot boundary detection, camera motion estimation, news story segmentation, high-level feature detection and video retrieval. Among these tasks, shot boundary and news story detection are events in video production; there is a number of high-level features or video queries that are related to visual events in the shot, such as "people-marching," and "one or more people entering or leaving a building." TRECVID and the LSCOM large scale ontology [98] fosters the definition, annotation, and detection of a large collection of semantic concepts and events.

Performance Evaluation of Tracing and Surveillance (PETS) has been held since 2000 for evaluating visual tracking and surveillance algorithms. This benchmark supplies multiview multicamera (up to four cameras) surveillance data [35] for detecting events such as left luggage in public spaces. Video Analysis and Content Extraction (VACE) is a government-funded program that aims developing novel algorithms and implementations for

automatic video content extraction, multimodal fusion and event understanding. VACE evaluation benchmarks systems for automated detection and tracking of scene objects such as faces, hands, humans, vehicles, and text in four primary video domains: broadcast news, meetings, surveillance, and UAV. It has been observed [88] that PETS and VACE have plenty of synergies in terms of evaluation goals, data domains and tasks, while the specifics in common tools, ground truth annotations, and metrics still need to be normalized.

Computer in the human interaction loop (CHIL) is an international consortium aiming to "realize computer services that are delivered to people in an implicit, indirect and unobtrusive way" [2]. The data are audio–visual streams recorded from surveillance, meetings or smart room application, the tasks include speaker localization, speaker tracking, multimodal interaction, etc. Augmented Multiparty Interaction (AMI) is an EU research consortium that "targets computer enhanced multimodal interaction in the context of meetings" [1] via test collections of instrumented meeting rooms: video footage from multiple cameras and microphones. The systems process the audio and visual track for a collection of both mid-level and high-level detection tasks such as: face detection, speaker identification, tracking focus of attention, and detecting participant influence. CLassification of Events, Activities and Relationships (CLEAR) [133] is a cross-campaign collaboration between VACE and CHIL, concerned with getting consensus and crossover on the evaluation of event classifications.

Several articles have argued the pros and cons of benchmark campaigns in multimedia information retrieval [128], [143]. The discussions apply to event detection and modeling in general, and we summarize them as follows. The advantages include: 1) Collect, prepare and distribute data, making results directly comparable across systems. 2) Create critical mass around common challenges so as to encourage donation of resources and collaboration. 3) Participating groups can learn from each other on the common grounds of the benchmark. This helps to accelerate the performance of new comers and helps bridge the gap between research ideas and practical systems. The commonly noted disadvantages include limiting the current and future problems being addressed in the community and reducing the room for diversity.

Setting up appropriate benchmarks for real-world events is a challenging task in itself, and the research community is making progress to define tasks beyond the components (i.e., objects, scenes, movements) of events. As a relatively new area with diverse problem definitions and solutions, it will benefit from shared tools and common platforms, such as feature extraction, tracking, and detection results of important objects such as face, car, people. Event evaluation also calls for new problems in new application domains, such as in user-generated content [14], and multimodal, multi-source, asynchronous event detection.

V. CONCLUSION

This paper presented a survey on multimedia event mining, covering aspects of event description, modeling, and analysis in multimedia content. This area has seen significant recent efforts from the research community, and this has led to new technologies and systems.

We study events as real-world occurrences that unfold over space and time. We introduced a framework for event description based on the five "W"s and one "H" in journalism. We showed how the six facets (*who*, *when*, *where*, *what*, *why*, and *how*) can be used for events description, and we also discussed how and why there is semantic variability. Event detection was presented as the process of mapping multimedia streams to event descriptions. We examined the five major components of event-modeling systems: target properties, data capture, feature representation, computational model, and applications. The target event properties, data capture, and application components defined the problem, while the feature representation and computational model made up the solution. A significant portion of this survey was devoted to the review of a range of existing and ongoing work on event detection. We identified several groups of event detection systems based on their problem setup and detection target. There are significant differences between the event detection problem and the detection of static objects and scenes. Event models produce structured output—the 5W1Hs, from semi-structured input—image/video/audio and metadata. Progress has been made in event detection components, such as background substraction, tracking, and the detection of single event facet, such as detecting faces, people (*who*), objects (*what*), location (*where*, indoor/outdoor), and time (*when*, day/night) by focusing on a tightly controlled environments or a short period of time where other facets remain constant. Finally, we also reviewed current benchmarks related to event detection.

This review can be put into perspective from three different aspects: 1) Solving the event detection problem from incomplete data is going to remain a significant challenge for many real-world applications. We focused on event analysis from existing archives and repositories, while problems addressing real-time capture and rich representation of media streams are likely to receive more attention. 2) The majority of the work reviewed here relied to a significant extent on visual information and less on other modalities such as sound, freeform text, or structured meta-tags. Multimodal fusion is an important challenge for multimedia research in general, and we are seeing more and more work on this direction. 3) New systems and algorithms are under active research in many organizations, while the systems reviewed in this paper will soon become a smaller subset of whole picture, our framework for thinking about the problems and solutions in the event mining space is likely to remain relevant.

Event mining is a vibrant area of research. Emerging topics of investigation include: 1) Distributed, multimodal multisource event modeling and detection, with immediate applications in analyzing surveillance and meeting recordings [8], [91]. 2) Event modeling and detection by explicitly modeling multiple facets [95], [167] also producing faceted annotation based on time and location proximity, as well as social interactions. 3) Extracting faceted attributes from unstructured or semi-structured data (image, exif metadata, or user-generated tags) [110]. 4) Event analysis that takes into account rich event capture and annotation [144]—how to use and disambiguate the faceted attributes when available and how to discern the relationships among events. 5) Systematically encode and estimate domain knowledge, and use this knowledge to improve recognition, save computation, as well as to help develop systems that will generalize across different data domains. ■

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable and constructive comments that helped improve the presentation of this paper substantially. This material is based upon work funded in part by the U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Government.

REFERENCES

[1] Augmented Multi-Person Interaction (AMI). [Online]. Available: http://www.amiproject.org

[2] Computer in the Human Interaction Loop. [Online]. Available: http://www.chil.server.de/servlet/is/101/

[3] Five Ws. [Online]. Available: http://www.en.wikipedia.org/wiki/Five_Ws

[4] Flickr Group: ACM Multimedia 2006. [Online]. Available: http://www.flickr.com/groups/acmmm2006/

[5] Popular Tags on Flickr Photo Sharing. [Online]. Available: http://www.flickr.com/photos/tags/

[6] The Oxford English Dictionary. [Online]. Available: http://www.oed.com/

[7] User-Generated Content. [Online]. Available: http://www.en.wikipedia.org/wiki/User-generated_content

[8] H. Aghajan and C. Wu, "Layered and collaborative gesture analysis in multi-camera networks," in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing (ICASSP)*, 2007.

[9] S. Ahern, M. Naaman, R. Nair, and J. H.-I. Yang, "World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections," in *Proc. JCDL '07: Conf. Digital Libraries*, New York, 2007, pp. 1–10.

[10] M. Ames and M. Naaman, "Why we tag: Motivations for annotation in mobile and online media," in *Proc. CHI '07: Conf. Human Factors in Computing Systems*, New York, 2007, pp. 971–980.

[11] P. Appan and H. Sundaram, "Networked multimedia event exploration," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, New York, 2004, pp. 40–47.

[12] O. Arandjelovic and A. Zisserman, "Automatic face recognition for film character retrieval in feature-length films," in *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition*, 2005, vol. 1.

[13] J. Assfalg, M. Bertini, C. Colombo, and A. D. Bimbo. (2002, Feb.). Semantic annotation of sports videos, *IEEE MultiMedia*. [Online]. *9(2)*, pp. 52–60. Available: http://www.computer.org:80/multimedia/mu2002/u2052abs.htm

[14] Associated Press, "Now starring on the web: YouTube," *Wired News*, Apr. 2006.

[15] P. Belhumeur, J. Hespanha, D. Kriegman *et al.*, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[16] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. A. Forsyth, "Names and faces in the news," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004, IEEE Computer Soc.

[17] D. Bordwell and K. Thompson, *Film Art: An Introduction*. New York: McGraw-Hill, 2001.

[18] J. S. Boreczky and L. A. Rowe. (1996). "Comparison of video shot boundary detection techniques," in *Proc. Storage Retrieval for Image and Video Databases (SPIE)*, pp. 170–179. [Online]. Available: citeseer.ist.psu.edu/boreczky96comparison.html

[19] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proc. Conf. Computer Vision and Pattern Recognition (CVPR '97)*, Washington, DC, 1997, p. 994, IEEE Computer Soc.

[20] R. Brunelli, O. Mich, and C. M. Modena. (1999, Jun.). A survey on the automatic indexing of video data, *J. Visual Commun. Image Representation*. [Online]. *10(2)*, pp. 78–112. Available: http://www.dx.doi.org/10.1006/jvci.1997.0404

[21] C. Cédras and M. Shah, "Motion-based recognition: A survey," *Image Vision Comput.*, vol. 13, no. 2, pp. 129–155, 1995.

[22] L. Chaisorn, T.-S. Chua, and C.-H. Lee, "A multi-modal approach to story segmentation for news video," *World Wide Web*, vol. 6, no. 2, pp. 187–208, 2003.

[23] S. F. Chang, T. Sikora, and A. Puri, "Overview of the MPEG-7 standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 688–695, Jun. 2001.

[24] S.-F. Chang, "The holy grail of content-based media analysis," *IEEE Multimedia Mag.*, vol. 9, no. 2, pp. 6–10, Apr. 2002.

[25] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui, and J. Luo, "Large-scale multimodal semantic concept detection for consumer video," in *Proc. ACM SIGMM Int. Workshop Multimedia Information Retrieval*, Germany, Sep. 2007.

[26] D. Chen, R. Malkin, and J. Yang, "Multimodal detection of human interaction events in a nursing home environment," in *Proc. 6th Int. Conf. Multimodal Interfaces*, New York, 2004, pp. 82–89.

[27] T.-S. Chua, S.-F. Chang, L. Chaisorn, and W. Hsu, "Story boundary detection in large broadcast news video archives techniques, experience and trends," *ACM Multimedia*, Oct. 2004, New York.

[28] B. Clarkson and A. Pentland, "Unsupervised clustering of ambulatory audio and video," in *Proc. Int. Conf. Acoustic, Speech and Signal Processing (ICASSP)*, 1999.

[29] M. Collins and Y. Singer, "Unsupervised models for named entity classification," in *Proc. Joint SIGDAT Conf. Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999, pp. 189–196.

[30] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox, "Temporal event clustering for digital photo collections," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 1, no. 3, pp. 269–288, 2005.

[31] J. W. Davis and A. F. Bobick, "The representation and recognition of action using temporal templates," *IEEE Computer Vision and Pattern Recognition (CVPR)*, pp. 928–934, 1997.

[32] M. Davis, "Editing out video editing," *IEEE Multimedia*, vol. 10, no. 2, pp. 54–64, Feb. 2003.

[33] M. Davis, S. King, N. Good, and R. Sarvas, "From context to content: Leveraging context to infer media metadata," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, New York, 2004, pp. 188–195.

[34] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoustical Soc. Amer.*, vol. 111, p. 1917, 2002.

[35] X. Desurmont, R. Sebbeb, F. Martina, C. Machya, and J.-F. Delaiglea, "Performance evaluation of frequent events detection systems," in *Proc. 9th IEEE Int. Workshop Performance Evaluation of Tracking and Surveillance*, 2006.

[36] A. Dey, *Providing Architectural Support for Context-Aware Applications*. Atlanta, GA: Georgia Inst. Technology, Nov. 2000.

[37] A. Dey and G. Abowd, "Towards a better understanding of context and context-awareness," in *Proc. CHI 2000 Workshop on What, Who, Where, When, and How of Context-Awareness*.

[38] P. Dourish, "What we talk about when we talk about context," *Personal Ubiquitous Computing*, vol. 8, no. 1, pp. 19–30, 2004.

[39] L.-Y. Duan, M. Xu, T.-S. Chua, Q. Tian, and C.-S. Xu, "A mid-level representation framework for semantic sports video

analysis," in *Proc. 11th ACM Int. Conf. Multimedia*, New York, 2003, pp. 33–44.

[40] S. Ebadollahi, L. Xie, S.-F. Chang, and J. R. Smith, "Visual event detection using multi-dimensional concept dynamics," in *Proc. Int. Conf. Multimedia and Expo (ICME)*, Toronto, Canada, Jul. 2006.

[41] A. Ekin, A. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. Image Processing*, vol. 12, no. 7, pp. 796–807, Jul. 2003.

[42] D. P. W. Ellis and K. Lee, "Accessing minimal-impact personal audio archives," *IEEE MultiMedia*, vol. 13, no. 4, pp. 30–38, Apr. 2006.

[43] H.-L. Eng, K.-A. Toh, A. H. Kam, J. Wang, and W.-Y. Yau, "An automatic drowning detection surveillance system for challenging outdoor pool environments," in *ICCV '03: Proc. 9th IEEE Int. Conf. Computer Vision*, Washington, DC, 2003, p. 532, IEEE Computer Soc.

[44] Z. Feng, X. Zhu, X. Wu, A. K. Elmagarmid, and L. Wu, "Video data mining: Semantic indexing and event detection from the association perspective," *IEEE Trans. Knowledge Data Eng.*, vol. 17, no. 5, pp. 665–677, 2005.

[45] M. Fleischman and D. Roy, "Situated models of meaning for sports video retrieval," in *Proc. HLT/NAACL*, Rochester, NY, 2007.

[46] M. Fleischman, P. Decamp, and D. Roy, "Mining temporal patterns of movement for video content classification," in *Proc. 8th ACM Int. Workshop Multimedia Information Retrieval*, New York, 2006, pp. 183–192.

[47] A. R. J. Francois, R. Nevatia, J. Hobbs, and R. C. Bolles, "Verl: An ontology framework for representing and annotating video events," *IEEE MultiMedia*, vol. 12, no. 4, pp. 76–86, Apr. 2005.

[48] U. Gargi, "Modeling and clustering of photo capture streams," in *Proc. 5th ACM SIGMM Int. Workshop Multimedia Information Retrieval*, New York, 2003, pp. 47–54.

[49] B. Gold and N. Morgan, *Speech and Audio Signal Processing*. New York: Wiley, 2000.

[50] A. Graham, H. Garcia-Molina, A. Paepcke, and T. Winograd, "Time as essence for photo browsing through personal digital libraries," in *Proc. 2nd ACM/IEEE-CS Joint Conf. Digital Libraries*, New York, 2002, pp. 326–335.

[51] S. Greenberg, "Context as a dynamic construct," *Human-Computer Interaction*, vol. 16, no. 2–4, pp. 257–268, 2001.

[52] A. Gupta and L. Davis, "Objects in action: An approach for combining action understanding and object perception," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Jun. 2007.

[53] A. Hakeem, Y. Sheikh, and M. Shah, "CASE E: A hierarchical event representation for the analysis of videos," in *Proc. 19th Nat. Conf. Artificial Intelligence (AAAI)*, San Jose, CA, 2004, pp. 263–268.

[54] R. Hamid, A. Johnson, S. Batta, A. Bobick, C. Isbell, and G. Coleman, "Detection and explanation of anomalous activities: Representing activities as bags of event n-grams," in *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition (CVPR'05)—Vol. 1*, Washington, DC, 2005, pp. 1031–1038.

[55] R. Hamid, S. Maddi, A. Bobick, and I. Essa, "Structure from statistics—Unsupervised activity analysis using suffix trees," in *Proc. Int. Conf. Computer Vision (ICCV)*, Rio de Janeiro, Brazil, Oct. 2007.

[56] L. Hardman, "Canonical processes of media production," in *Proc. ACM Workshop Multimedia for Human Communication*, New York, 2005, pp. 1–6.

[57] T. Hastie, R. Tibshirani, J. Friedman *et al.*, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2001.

[58] A. G. Hauptmann and M. J. Witbrock. (1998). "Story segmentation and detection of commercials in broadcast news video," in *Advances in Digital Libraries*, pp. 168–179. [Online]. Available: http://www.citeseer.ist.psu.edu/hauptmann98story.html

[59] E. Hjelmas and B. Low, "Face detection: A survey," *Computer Vision and Image Understanding*, vol. 83, no. 3, pp. 236–274, 2001.

[60] S. Hongeng, F. Bremond, and R. Nevatia, "Representation and optimal recognition of human activities," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2000, vol. 1, pp. 818–825.

[61] S. Hongeng and R. Nevatia, "Multi-agent event recognition," in *Proc. IEEE Int. Conf. Computer Vision (ICCV'01)*, 2001, vol. 2, pp. 84–91.

[62] B. Horn, *Robot Vision*. New York: McGraw-Hill, 1986.

[63] W. Hsu, S.-F. Chang, C.-W. Huang, L. Kennedy, C.-Y. Lin, and G. Iyengar, "Discovery and fusion of salient multi-modal features towards news story segmentation," in *Proc. Symp. Electronic Imaging: Science Technology—SPIE Storage and Retrieval of Image/Video Database*, San Jose, CA, Jan. 2004.

[64] J. Huang, S. Ravi Kumar, M. Mitra, W. Zhu, and R. Zabih, "Spatial color indexing and applications," *Int. J. Computer Vision*, vol. 35, no. 3, pp. 245–268, 1999.

[65] Y. A. Ivanov and A. F. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 852–872, Aug. 2000.

[66] T. Jaakkola, M. Diekhans, and D. Haussler, "Using the Fisher kernel method to detect remote protein homologies," in *Proc. 7th Int. Conf. Intelligent Systems for Molecular Biology*, 1999, pp. 149–158.

[67] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[68] T. Jebara, *Machine Learning: Discriminative and Generative*. New York: Kluwer, 2004.

[69] JEITA, *Exchangeable image file format for digital still cameras: Exif Version 2.2*, Apr. 2002.

[70] F. Jelinek, *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press, 1997.

[71] S.-W. Joo and R. Chellappa, "Attribute grammar-based event recognition and anomaly detection," in *Proc. Conf. Computer Vision and Pattern Recognition Workshop*, Washington, DC, 2006, p. 107.

[72] D. Jurafsky and J. Martin, *Speech and Language Processing*. Upper Saddle River, NJ: Prentice Hall, 2000.

[73] J. Kender and B. Yeo, "Video scene segmentation via continuous video coherence," in *Proc. IEEE Computer Soc. Conf.*, 1998, pp. 367–373.

[74] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, "How Flickr helps us make sense of the world: Context and content in community-contributed media collections," in *Proc. 15th Int. Conf. Multimedia*, New York, 2007, pp. 631–640.

[75] D. Kirk, A. Sellen, R. Harper, and K. Wood, "Understanding videowork," in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, New York, 2007, pp. 61–70.

[76] D. Kirk, A. Sellen, C. Rother, and K. Wood, "Understanding photowork," in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, New York, 2006, pp. 761–770.

[77] Y. LeCun and Y. Bengio, *Convolutional Networks for Images, Speech, and Time Series*. Cambridge, MA: MIT Press, 1998, pp. 255–258.

[78] M. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Computing, Communications, Applications (TOMCCAP)*, vol. 2, no. 1, pp. 1–19, 2006.

[79] B. Li and M. I. Sezan, "Event detection and summarization in sports video," in *CBAIVL '01: Proc. IEEE Workshop Content-based Access of Image and Video Libraries*, Washington, DC, 2001, p. 132.

[80] X. Li, D. Wang, J. Li, and B. Zhang, "Video search in concept subspace: A text-like paradigm," in *Proc. Int. Conf. Image and Video Retrieval (CIVR)*, Amsterdam, The Netherlands, Jul. 2007.

[81] J.-H. Lim, Q. Tian, and P. Mulhem. (2003, Apr.). Home photo content modeling for personalized event-based retrieval, *IEEE MultiMedia*. [Online]. *10(4)*, pp. 28–37. Available: http://csdl.computer.org/comp/mags/mu/2003/04/u4028abs.htm

[82] Z. Liu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," *J. VLSI Signal Processing*, vol. 20, no. 1, pp. 61–79, 1998.

[83] Z. Liu, D. Gibbon, E. Zavesky, B. Shahraray, and P. Haffner, "A fast, comprehensive shot boundary determination system," in *Proc. IEEE Int. Conf. Multimedia and Expo*, Beijing, China, Jul. 2007, pp. 1487–1490.

[84] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text classification using string kernels," *J. Machine Learning Res.*, vol. 2, pp. 419–444, 2002.

[85] A. C. Loui and A. E. Savakis, "Automated event clustering and quality screening of consumer pictures for digital albuming," *IEEE Trans. Multimedia*, vol. 5, no. 3, pp. 390–402, Mar. 2003.

[86] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[87] A. Mani and H. Sundaram, "Modeling user context with applications to media retrieval," *Multimedia Systems J.*, vol. 12, no. 4, pp. 339–353, 2007.

[88] V. Manohar, M. Boonstra, V. Korzhova, P. Soundararajan, D. Goldgof, R. Kasturi, S. Prasad, H. Raju, R. Bowers, and J. Garofolo, "Pets vs. vace evaluation programs: A comparative study," in *Proc. 9th IEEE Int. Workshop Performance Evaluation of Tracking and Surveillance*, 2006.

[89] J. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoustics*, vol. 20, no. 5, pp. 367–377, 1972.

[90] W. Marslen-Wilson and A. Welsh, *Processing Interactions and Lexical Access During Word Recognition in Continuous Speech*, Dept. Behavioral Sciences, Univ. Chicago, 1977.

[91] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 305–317, Mar. 2005.

[92] G. Medioni, I. Cohen, F. Brémond, S. Hongeng, and R. Nevatia, "Event detection and analysis from video streams," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 8, pp. 873–889, Aug. 2001.

[93] E. T. Mueller, *Event Calculus*. London, U.K.: Elsevier, 2007.

[94] M. Naaman, S. Harada, Q. Wang, H. Garcia-Molina, and A. Paepcke, "Context data in geo-referenced digital photo collections," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, New York, 2004, pp. 196–203.

[95] M. Naaman, R. B. Yeh, H. Garcia-Molina, and A. Paepcke, "Leveraging context to resolve identity in photo albums," in *Proc. 5th ACM/IEEE-CS Joint Conf. Digital Libraries*, New York, 2005, pp. 178–187.

[96] F. Nack, "Capture and transfer of metadata during video production," in *Proc. ACM Workshop Multimedia for Human Communication*, New York, 2005, pp. 17–20.

[97] M. R. Naphade, T. Kristjansson, B. Frey, and T. S. Huang, "Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems," in *Proc. IEEE Int. Conf. Image Processing (ICIP'98)*, Oct. 1998, vol. 3, pp. 536–540.

[98] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE MultiMedia*, vol. 13, no. 3, pp. 86–91, Mar. 2006.

[99] A. Natsev, A. Haubold, J. Tesic, R. Yan, and L. Xie, "Semantic concept-based query expansion and re-ranking for multimedia retrieval," in *Proc. ACM Multimedia*, Augsburg, Germany, Sep. 2007.

[100] S.-Y. Neo, J. Zhao, M.-Y. Kan, and T.-S. Chua, "Video retrieval using high level features: Exploiting query matching and confidence-based weighting," in *Proc. Image and Video Retrieval, 5th Int. Conf.*, Tempe, AZ, Jul. 2006, pp. 143–152.

[101] R. Nevatia, T. Zhao, and S. Hongeng, "Hierarchical language-based representation of events in video streams," in *Proc. IEEE Workshop Event Mining (EVENT '03)*, 2003.

[102] N. M. Oliver, B. Rosario, and A. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 8, Aug. 2000.

[103] V. Parameswaran and R. Chellappa, "View invariants for human action recognition," in *IEEE Computer Vision Pattern Recognition*, 2003, pp. 613–619.

[104] V. A. Petrushin, "Mining rare and frequent events in multi-camera surveillance video using self-organizing maps," in *Proc. 11th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining*, New York, 2005, pp. 794–800.

[105] J. C. Platt, M. Czerwinski, and B. A. Field. (2003). "PhotoTOC: Automatic clustering for browsing personal photographs," in *Proc. 4th IEEE Pacific Rim Conf. Multimedia*. [Online]. Available: citeseer.ist.psu.edu/czerwinski02phototoc.html

[106] J. Pomerol and P. Brézillon, "About some relationships between knowledge and context," in *Modeling and Using Context (CONTEXT-01)*, vol. 1688, *Lecture Notes in Computer Science*. New York: Springer Verlag, 2001, pp. 461–464.

[107] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall, 1993.

[108] R. Radhakrishnan, A. Divakaran, and Z. Xiong, "A time series clustering based framework for multimedia mining and summarization using audio features," in *Proc. 6th ACM SIGMM Int. Workshop Multimedia Information Retrieval*, New York, 2004, pp. 157–164.

[109] C. Rao, A. Yilmaz, and M. Shah, "View-invariant representation and recognition of actions," *Int. J. Computer Vision*, vol. 50, no. 2, pp. 203–226, 2002.

[110] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from Flickr tags," in *Proc. Int. ACM SIGIR Conf. Res. Development in Information Retrieval*, New York, 2007.

[111] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[112] K. Rodden and K. R. Wood, "How do people manage their digital photographs?" in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, New York, 2003, pp. 409–416.

[113] R. Rosenfeld, "A whole sentence maximum entropy language model," in *Proc. IEEE Workshop Automatic Speech Recognition and Understanding*, 1997, pp. 230–237.

[114] W. Roush. (2007, Mar.). Tr10: Peering into video's future, *Technol. Rev.* [Online]. Available: http://www.technologyreview.com/Infotech/18284/?a=f

[115] D. Roy, R. Patel, P. DeCamp, R. Kubat, M. Fleischman, B. Roy, N. Mavridis, S. Tellex, A. Salata, J. Guinness *et al.*, "The human speechome project," *Cognitive Sci.*, 2006.

[116] Y. Rubner, C. Tomasi, and L. Guibas, "The Earth mover's distance as a metric for image retrieval," *Int. J. Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[117] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in *Proc. 8th ACM Int. Conf. Multimedia*, 2000, pp. 105–115.

[118] Y. Rui, T. Huang, and S. Chang, "Image retrieval: Current techniques, promising directions and open issues," *J. Visual Commun. Image Representation*, vol. 10, no. 4, pp. 39–62, 1999.

[119] S. Satoh, Y. Nakamura, and T. Kanade, "Name-It: Naming and detecting faces in news videos," *IEEE Multimedia*, vol. 6, no. 1, pp. 22–35, Jan. 1999.

[120] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP '97)*, Washington, DC, 1997, vol. 2, p. 1331.

[121] J. Schlenzig, E. Hunter, and R. Jain. (1994). "Recursive identification of gesture inputs using hidden Markov models," in *Proc. 2nd IEEE Workshop Applications of Computer Vision*, pp. 187–194. [Online]. Available: http://vision.ucsd.edu/schlenz/

[122] M. Shanahan, "The event calculus explained," in *Artificial Intelligence Today: Recent Trends and Developments*, 1999

[123] L. Shapiro and G. Stockman, *Computer Vision*. Upper Saddle River, NJ: Prentice Hall, 2001.

[124] S. Sharff, *The Elements of Cinema: Toward a Theory of Cinesthetic Impact*. New York: Columbia Univ. Press, 1982.

[125] B. Shevade, H. Sundaram, and L. Xie, "Modeling personal and social network context for event annotation in images," in *Proc. 6th ACM/IEEE-CS Joint Conf. Digital Libraries*, 2007.

[126] Y. Shi, Y. Huang, D. Minnen, A. Bobick, and I. Essa, "Propagation networks for recognition of partially ordered sequential action," in *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition (CVPR 2004)*, Washington, DC, Jun. 2004, pp. II862–II870.

[127] J. Slate. (2002, Dec.). *A Guide to Metadata*. [Online]. http://www.cetis.ac.uk

[128] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVID," in *Proc. 8th ACM Int. Workshop Multimedia Information Retrieval*, New York, 2006, pp. 321–330.

[129] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[130] C. G. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring, "Adding semantics to detectors for video retrieval," *IEEE Trans. Multimedia*, vol. 14, no. 8, Aug. 2007.

[131] C. G. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools Applic.*, vol. 25, no. 1, pp. 5–35, 2005.

[132] T. Starner, A. Pentland, and J. Weaver, "Real-time American sign language recognition using desk and wearable computer based video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1371–1375, Dec. 1998.

[133] R. Stiefelhagen and R. Bowers, *Classification of Events, Activities and Relationships (CLEAR) Evaluation and Workshop*, 2006–2007. [Online]. Available: http://www.clear-evaluation.org/

[134] H. Sundaram and S.-F. Chang, "Determining computable scenes in films and their structures using audio-visual memory models," in *Proc. 8th ACM Int. Conf. Multimedia*, New York, 2000, pp. 95–104.

[135] M. Swain and D. Ballard, "Color indexing," *Int. J. Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[136] T. F. Syeda-Mahmood, M. A. O. Vasilescu, and S. Sethi, "Recognizing action events from multiple viewpoints," in *Proc. IEEE CVPR Workshop Detection Recognition of Events in Video*, 2001, p. 64.

[137] C. Teh and R. Chin, "On image analysis by the methods of moments," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 10, no. 4, pp. 496–513, Apr. 1988.

[138] Nat. Inst. Standards and Technology (NIST), *Text Retrieval Conference (TREC)*. [Online]. Available: http://trec.nist.gov/

[139] Nat. Inst. Standards and Technology (NIST), *TREC Video Retrieval Evaluation*, 2001–2007. [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid/

[140] P. Turaga, A. Veeraraghavan, and R. Chellappa, "Mining videos for events using a cascade of dynamical systems: From videos to verbs," in *Proc. IEEE Conf.*

Computer Vision and Pattern Recognition (CVPR), Jun. 2007.

[141] S. Velipasalar, L. M. Brown, and A. Hampapur, "Specifying, interpreting and detecting high-level, spatio-temporal composite events in single and multi-camera systems," in *Proc. Conf. Computer Vision and Pattern Recognition Workshop*, Washington, DC, 2006, p. 110, IEEE Computer Soc.

[142] J. Vendrig and M. Worring. (2003, Mar.). Interactive adaptive movie annotation. *IEEE MultiMedia*. [Online]. *10(3)*, pp. 30–37. Available: http://csdl.computer.org/comp/mags/mu/2003/03/u3030abs.htm

[143] J. Z. Wang, N. Boujemaa, A. D. Bimbo, D. Geman, A. G. Hauptmann, and J. Tesi, "Diversity in multimedia information retrieval research," in *Proc. 8th ACM Int. Workshop Multimedia information Retrieval*, New York, 2006, pp. 5–12.

[144] U. Westermann and R. Jain, "Toward a common event model for multimedia applications," *IEEE MultiMedia*, vol. 14, no. 1, pp. 19–29, Jan. 2007.

[145] T. Winograd, "Architectures for context," *Human-Computer Interaction*, vol. 16, no. 2–4, pp. 401–419, 2001.

[146] W. Wolf, B. Ozer, and T. Lv, "Smart cameras as embedded systems," *Computer*, vol. 35, no. 9, pp. 48–53, 2002.

[147] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden Markov models," in *Proc. Int. Conf. Acoustic, Speech and Signal Processing (ICASSP)*, Orlando, FL, 2002.

[148] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, *UnSupervised Mining of Statistical Temporal Structures in Video*. New York: Kluwer, 2003.

[149] L. Xie, L. Kennedy, S.-F. Chang, A. Divakaran, H. Sun, and C.-Y. Lin, "Discover meaningful multimedia patterns with audio-visual concepts and associated text," in *Proc. Int. Conf. Image Processing (ICIP)*, Oct. 2004.

[150] L. Xie, D. Xu, S. Ebadollahi, K. Scheinberg, S.-F. Chang, and J. R. Smith, "Pattern mining in visual concept streams," in *Proc. 40th Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Oct. 2006.

[151] L. Xie and R. Yan, "Extracting semantics from multimedia content: Challenges and solutions," in *Multimedia Content Analysis*. New York: Springer, 2008.

[152] E. P. Xing, R. Yan, and A. G. Hauptmann, "Mining associated text and images using dual-wing harmoniums," in *Proc. Uncertainty Artifical Intelligence (UAI)'05*, 2005.

[153] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. Huang, "Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework," in *Proc. Int. Conf. Multimedia and Expo ICME'03*, 2003, vol. 3.

[154] D. Xu and S.-F. Chang, "Visual event recognition in news video using kernel methods with multi-level temporal alignment," in *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition*, Minneapolis, MN, Jun. 2007.

[155] R. Yan and A. G. Hauptmann, "A review of text and image retrieval approaches for broadcast news video," *Inf. Retr.*, vol. 10, no. 4–5, pp. 445–484, 2007.

[156] J. Yang and A. Hauptmann, "Annotating news video with locations," in *Proc. Int. Conf. Image and Video Retrieval*, 2006.

[157] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, p. 13, 2006.

[158] X. Yu, C. Xu, H. W. Leong, Q. Tian, Q. Tang, and K. W. Wan, "Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video," in *Proc. 11th ACM Int. Conf. Multimedia*, New York, 2003, pp. 11–20.

[159] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B. Zhang, "A formal study of shot boundary detection," *IEEE Trans. Circuits Systems Video Technol.*, vol. 17, no. 2, pp. 168–186, Jun. 2007.

[160] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," in *Proc. CVPR*, 2001, pp. 123–130, IEEE Computer Society.

[161] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Semi-supervised adapted HMMs for unusual event detection," in *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition (CVPR'05)*, Washington, DC, 2005, vol. 1, pp. 611–618.

[162] H. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Syst.*, vol. 1, no. 1, pp. 10–28, 1993.

[163] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[164] H. Zhong, J. Shi, and M. Visontai, "Detecting unusual activity in video," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004, pp. 819–826.

[165] H. Zhou and D. Kimber, "Unusual event detection via multi-camera video mining," in *Proc. 18th Int. Conf. Pattern Recognition*, Washington, DC, 2006, pp. 1161–1166, IEEE Computer Soc.

[166] X. Zhu, "Semi-Supervised learning literature survey," Computer Sciences, Univ. Wisconsin-Madison, Tech. Rep. 1530, 2005.

[167] A. Zunjarwad, H. Sundaram, and L. Xie, "Contextual wisdom: Social relations and correlations for multimedia event annotation," in *Proc. 15th Int. Conf. Multimedia*, New York, 2007, pp. 615–624.

ABOUT THE AUTHORS

Lexing Xie received the B.S. degree from Tsinghua University, Beijing, China, in 2000, and the M.S. and Ph.D. degrees from Columbia University, in 2002 and 2005, respectively, all in electrical engineering.

She is a Research Staff Member at the IBM T.J. Watson Research Center, Hawthorne, NY. Her general research interests have included multimedia signal processing, content analysis, data mining, and machine learning. She has specifically worked on mining events and temporal patterns in multimedia, recognition, and search on large media collections and collaborative media annotation.

Dr. Xie has received several best student paper awards: 2007 ACM/IEEE Joint Conference on Digital Library (JCDL) on image annotation in social networks, IEEE International Conference on Image Processing (ICIP) in 2004 on discovering meaningful temporal patterns from video; ACM Multimedia 2005 on geometric features for distinguishing photo from computer graphics; ACM Multimedia 2002 on generating coherent audio-visual skims. She was the sole winner of 2005 IBM Research Josef Raviv Memorial Postdoc fellowship in computer science and engineering.

Hari Sundaram received the B.Tech. degree from the Indian Institute of Technology, Delhi, in 1993, the M.S. degree from the State University of New York, Stony Brook, in 1995, and the Ph.D. degree from Columbia University, New York, in 2002, all in electrical engineering.

He is currently an Assistant Professor at Arizona State University, Tempe. This is a joint appointment with the Department of Computer Science and the Arts Media and Engineering program. His research group works on developing computational models and systems for situated communication. Specific projects include context models for action, resource adaptation, interaction architectures, communication patterns in media sharing social networks, collaborative annotation, as well analysis of online communities.

Dr. Sundaram has won several awards—the best student paper award at JCDL 2007, the best ACM Multimedia demo award in 2006, the best student paper award at ACM conference on Multimedia 2002, and the 2002 Eliahu I. Jury Award for best Ph.D. dissertation. He has also received a best paper award on video retrieval, from IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, for the year 1998. He is an Associate Editor for *ACM Transactions on Multimedia Computing, Communications and Applications (TOMCCAP)*, as well as the *IEEE Signal Processing Magazine*.

Murray Campbell received the B.Sc. and M.Sc. degrees in computing science from the University of Alberta, Canada, and the Ph.D. degree in computer science from Carnegie Mellon University, Pittsburgh, PA, in 1987.

He has been with IBM Research since 1989. Currently, he is a Senior Manager in the Mathematical Sciences Department, IBM T.J. Watson Research Center, Yorktown Heights, NY. His group focuses primarily on the application of optimization, forecasting, probabilistic analysis, and expertise sharing to problems in business analytics and workforce management. His research interests also include the application of surveillance and early warning approaches to areas such as public health and petroleum production. Previously, he was a member of the team that developed Deep Blue, which in 1997 became the first computer to defeat the reigning world chess champion in a regulation match.

Dr. Campbell was awarded the Fredkin Prize and the Allen Newell Research Excellence Medal.